MESOSCALE NETWORK ANALYSIS
COMMUNITY DETECTION, CORE-PERIPHERY ANALYSIS

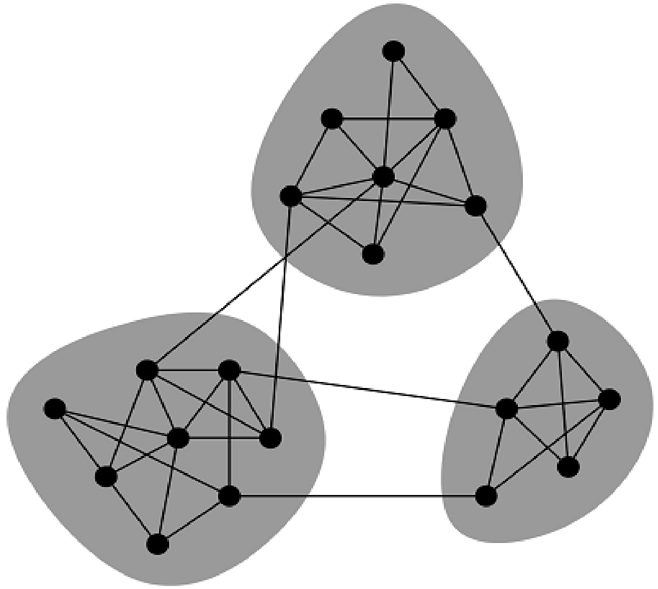
Carlo PICCARDI

DEIB - Department of Electronics, Information and Bioengineering
Politecnico di Milano, Italy

email carlo.piccardi@polimi.it
https://piccardi.faculty.polimi.it

COMMUNITY DETECTION

COMMUNITY ANALYSIS



Community analysis (or graph clustering) is aimed at revealing groups of nodes (communities) with dense intra- but sparse inter-community connections.

Important applications in biology, social networking, economics and finance, telecom, computer science, correlation networks, ...

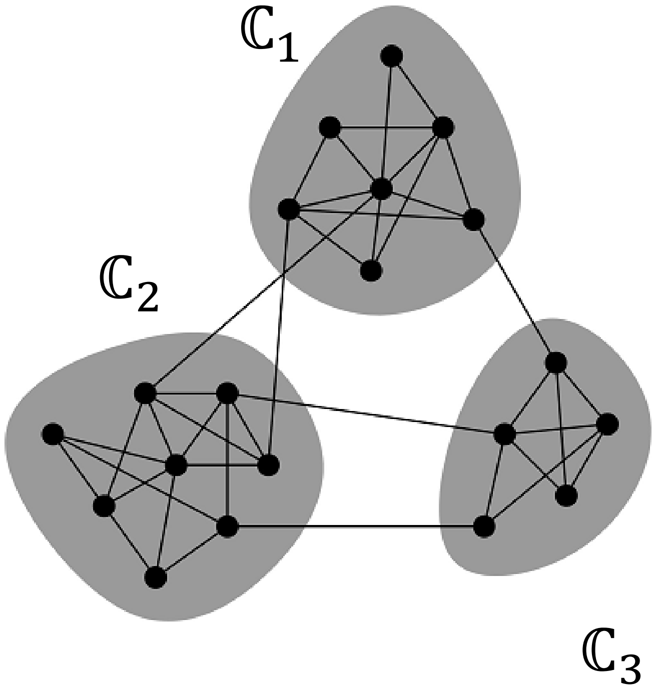
Plenty of methods *[Fortunato, Phys. Rep., 2010]*: "traditional" graph theory, betweenness-based, modularity-based, "dynamical" methods, statistical inference, ...

- PROBLEM 1 (CONCEPTUAL): How to rigorously define a community? That is, when a sub-network can be considered to be a significant cluster?

- PROBLEM 2 (TECHNICAL): For a N-node network, the "best" partition must be sought for in a set growing faster than $\exp(N)$: effective algorithms are needed.

  n. of partitions = B_N = Bell number (e.g., $B_5 = 52, B_{10} = 115975, B_{20} > 5 \times 10^{13}, \ldots$)

Modularity optimization *[Newman, PNAS, 2006]*



$\mathbb{C}_c$ is a set of nodes (a "candidate" community) and
$\mathbb{P}_q = \{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_q\}$ is a partition.

The modularity Q quantifies to what extent the intra-/inter-community link densities are anomalous in comparison to chance (i.e., to their expected value: "null model").

$Q =$ *(fraction of links internal to communities) - (expected fraction of such links)*

$$= \frac{1}{2L} \sum_{h=1,2,\ldots,q} \sum_{ij \in C_h} \left[a_{ij} - \frac{k_i k_j}{2L} \right]$$

A large value of Q (i.e. $Q \to 1$) typically reveals a significant community structure.

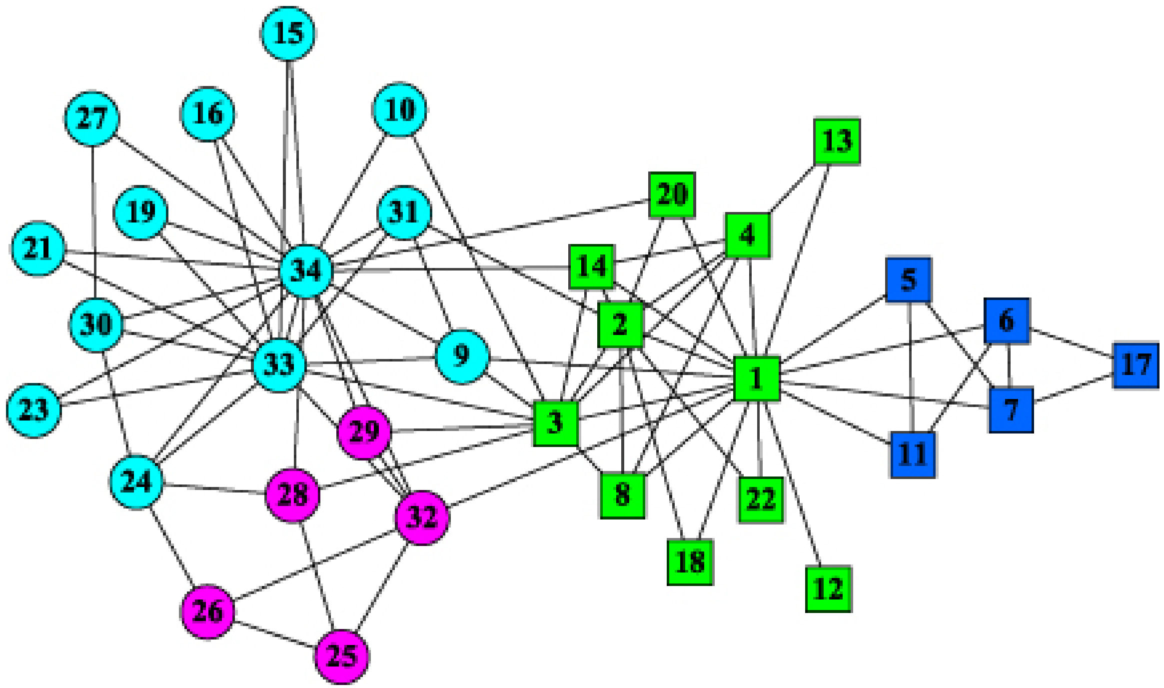
[Modularity optimization]: find the partition that maximizes Q.

$$Q = \frac{1}{2L} \sum_{h=1,2,\ldots,q} \sum_{ij \in C_h} \left[a_{ij} - \frac{k_i k_j}{2L} \right] = \sum_{h=1,2,\ldots,q} \left[\frac{L_h}{L} - \left(\frac{k_h}{2L} \right)^2 \right]$$

where L_h is the n. of links internal to C_h, and $k_h = \sum_{i \in C_h} k_i$ is the total degree of C_h.

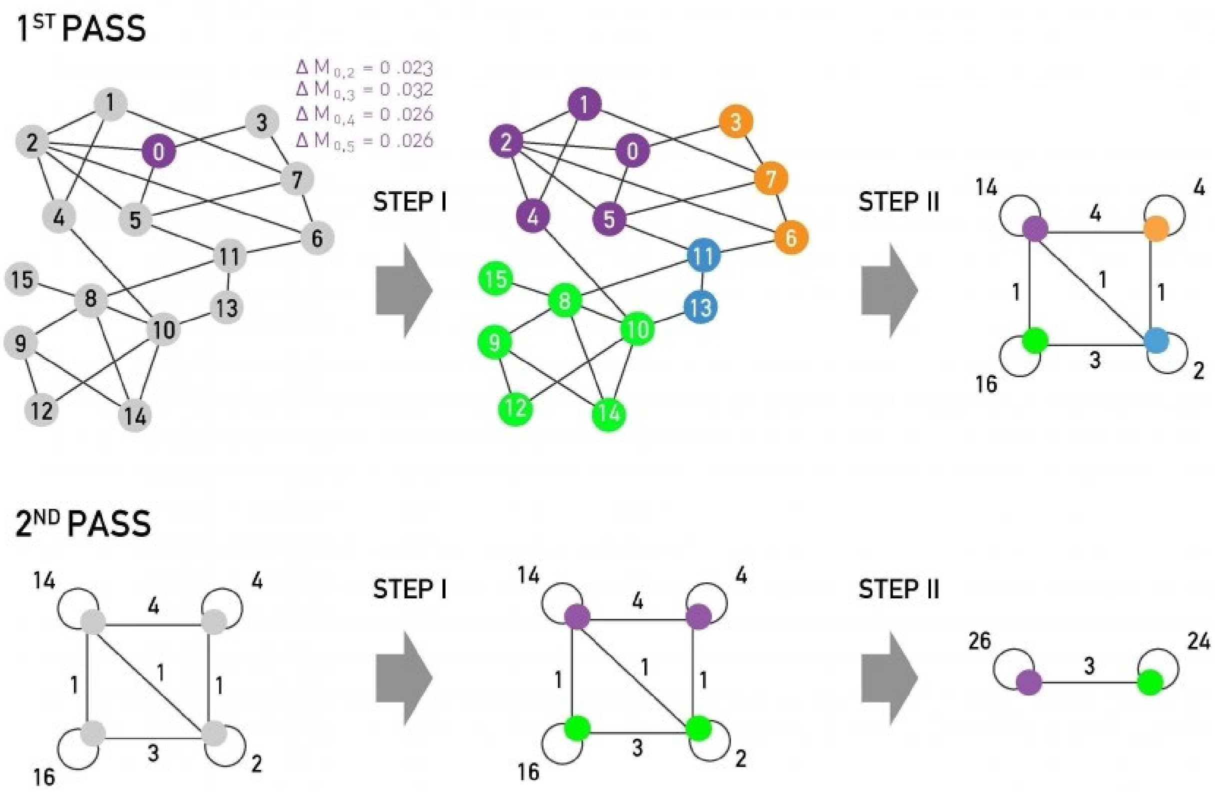
Example: Zachary's "karate club" social network

- The 4-community partition has the maximal modularity ($Q = 0.417$) among all partitions.
- E.g., the 2-community partition {darkblue ∪ green},{lightblue ∪ pink} has $Q = 0.371$.
- Predictive capability: the actual (historical) fission of the "karate club" is the 2-community partition squares/circles.



An exact solution to modularity optimization is practically unfeasible.

Many suboptimal algorithms are available: the most popular/fast is the Louvain method *[Blondel et al 2008]* ($O(n)$ in typical cases).



[from Barabasi 2016]

Each *pass* is composed of:

Step I: increase modularity by moving nodes to adjacent communities (try all nodes, move only if $\Delta Q > 0$ – formula for efficient computation of ΔQ !).

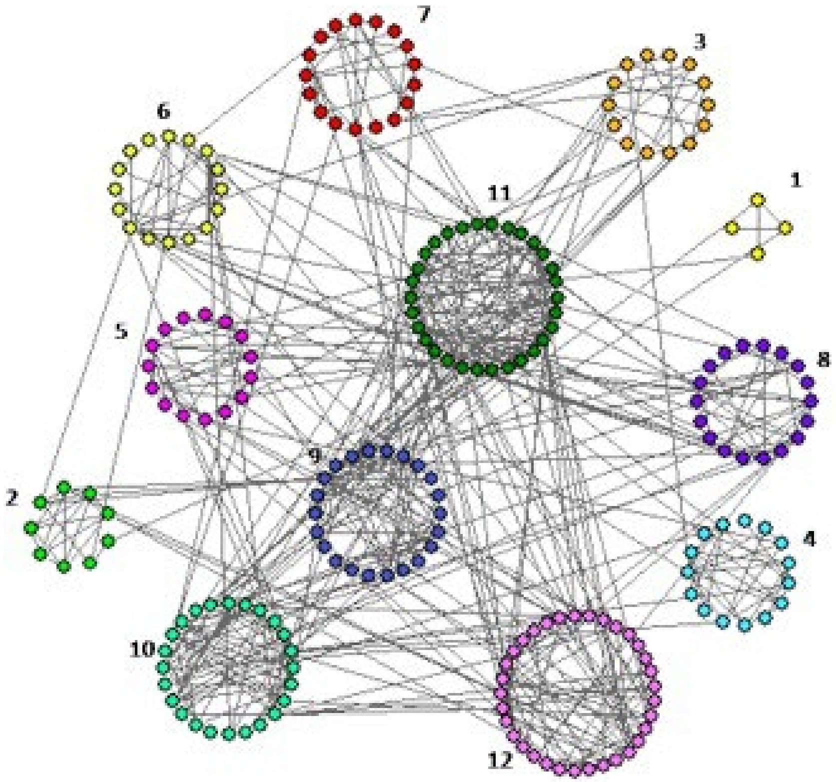
Step II: build a meta-network by aggregating nodes of the same community.

Repeat a new pass on the latest meta-network.

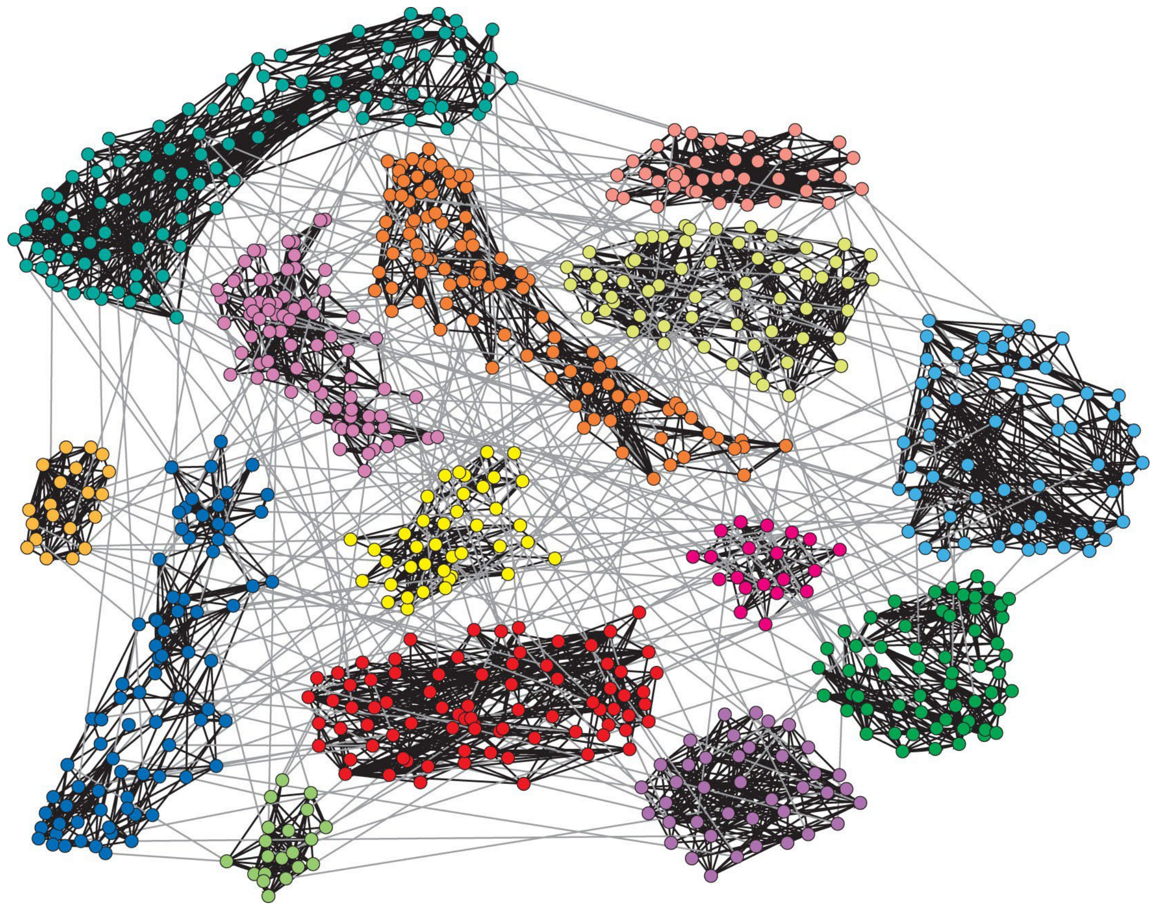
Stop when no further improvement of Q is possible.

Leiden method *[Traag et al 2019]* improves the method by solving some issues.

Applying the Louvain method to medium-scale networks:



Board interlocking of Italian companies
[Piccardi et al, PhysA, 2010]



Human brain neuronal system
[Russo et al, SciRep, 2014]

Extensions to the max-modularity method:

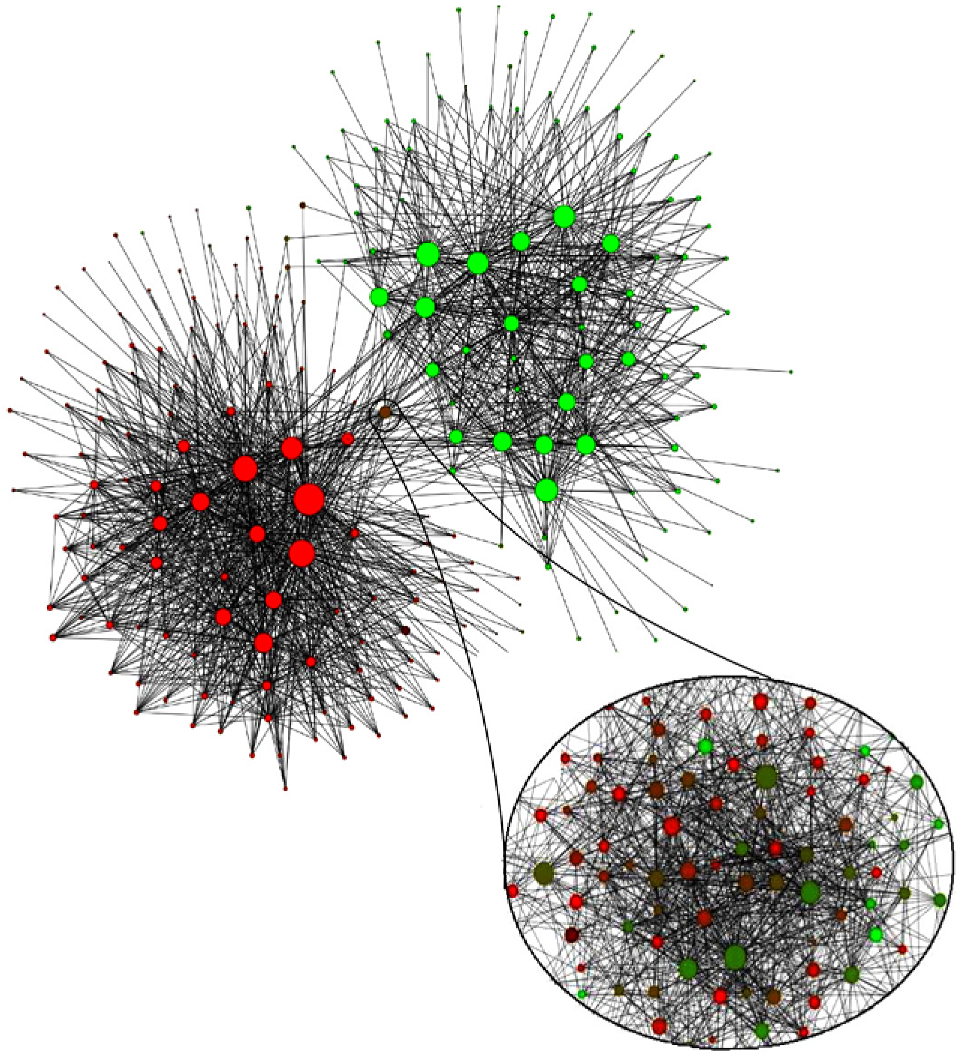
- directed and weighted networks
- overlapping communities
- hierarchical methods for very large networks
- ...

A few drawbacks:

- need to check "a posteriori" the quality of the resulting partition (*any network has a max-modularity partition!)* – see next page

- lacks to quantify the individual quality of each community – see next page

- since it forces a partition, it might miss to highlight a single strong community to favour the global optimization



- in very large networks, very small communities are missed ("resolution limit")

How to measure the quality of partitions/communities?

for partitions (`partition_quality` in NetworkX):

- Coverage ($0 \le \mathcal{C} \le 1$): the ratio of the number of intra-community edges to the total number of edges in the graph *(=the first term of the modularity Q)*.

$$\mathcal{C} = \sum_{h=1,2,\dots,q} \sum_{ij \in C_h} \frac{a_{ij}}{2L} = \sum_{h=1,2,\dots,q} \frac{L_h}{L}$$

- Performance ($0 \le \mathcal{P} \le 1$): the number of intra-community edges plus inter-community non-edges divided by the total number of potential edges.

$$\mathcal{P} = \frac{\left|\{(i,j) \text{ in the same community and } a_{ij} = 1\}\right| + \left|\{(i,j) \text{ in different communities and } a_{ij} = 0\}\right|}{N(N-1)/2}$$

for individual communities

- Persistence probability ($0 \le \alpha_{C_h} \le 1$): the ratio of the sum of the internal degrees of the nodes of C_h to the sum of the (total) degrees (*more follows to justify the name…*)

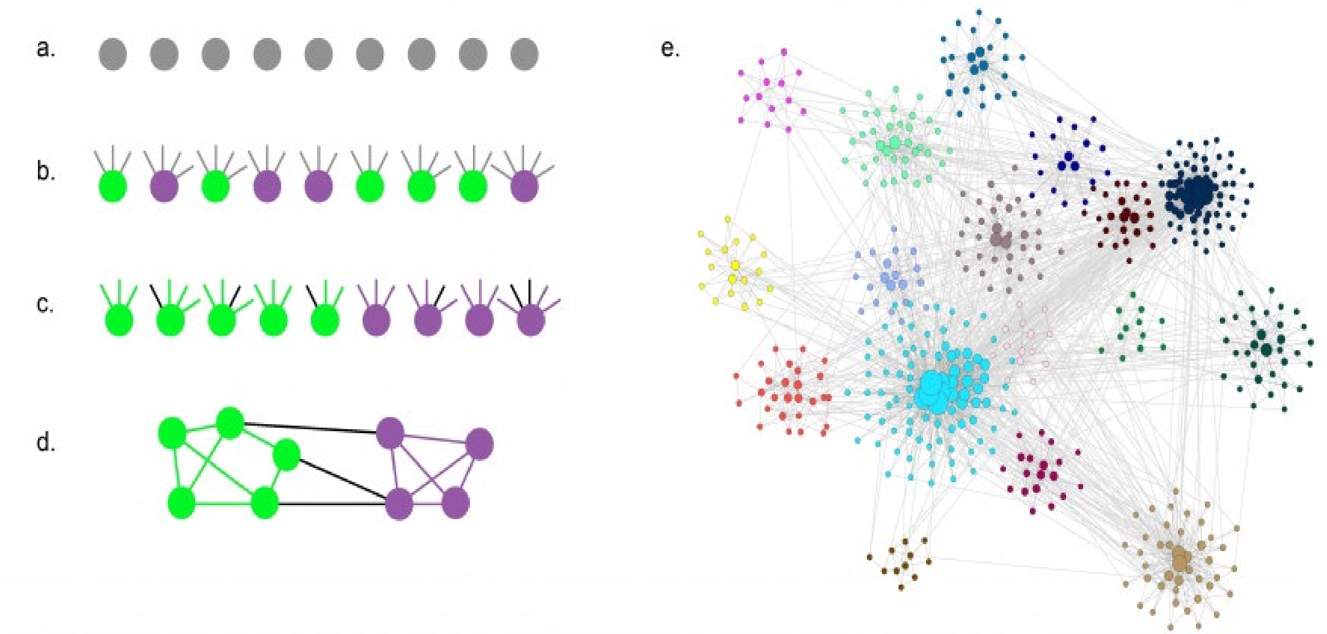
$$\alpha_{C_h} = \frac{\sum_{i \in C_h} k_i^{int}}{\sum_{i \in C_h} k_i} = \frac{\sum_{ij \in C_h} a_{ij}}{\sum_{i \in C_h} \sum_{j \in \{1,2,\dots,N\}} a_{ij}}$$

…more on network models: LFR (Lancichinetti-Fortunato-Radicchi) model

It is a block model creating a network with "realistic" planted community structure:

- with heterogeneous node degrees $P(k) \approx k^{-\gamma}$
- with heterogeneous community sizes $P(N_c) \approx N_c^{-\delta}$
- with tunable intra-/inter-community connectivity $(0 < \mu < 1)$

(a) Start with N isolated nodes.
(b) Select community sizes and assign each node to a community.
(c) For each node i select the degree k_i: the fraction μk_i will connect outside the community, the rest $(1 - \mu)k_i$ inside.
(d) Randomly connect intra- and inter-community links.



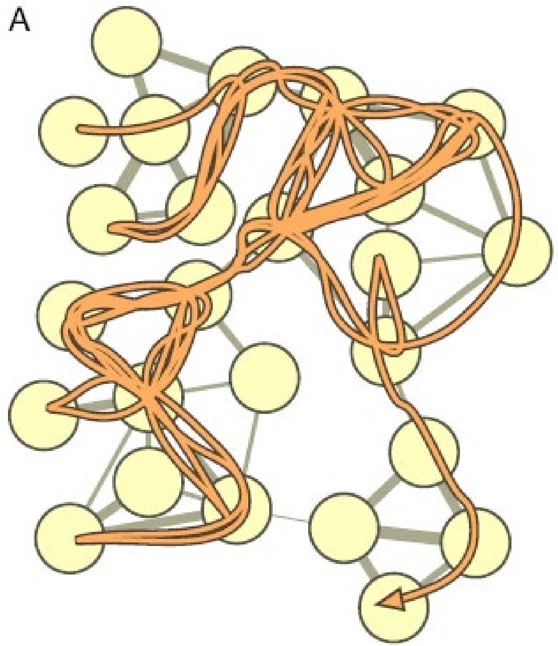
from Barabasi, 2016

Extensions to weighted and directed networks, and to overlapping communities.

Finding communities by means of random walkers

Given a community, links internally directed are many more (and/or with much larger weights) than links towards the rest of the network.

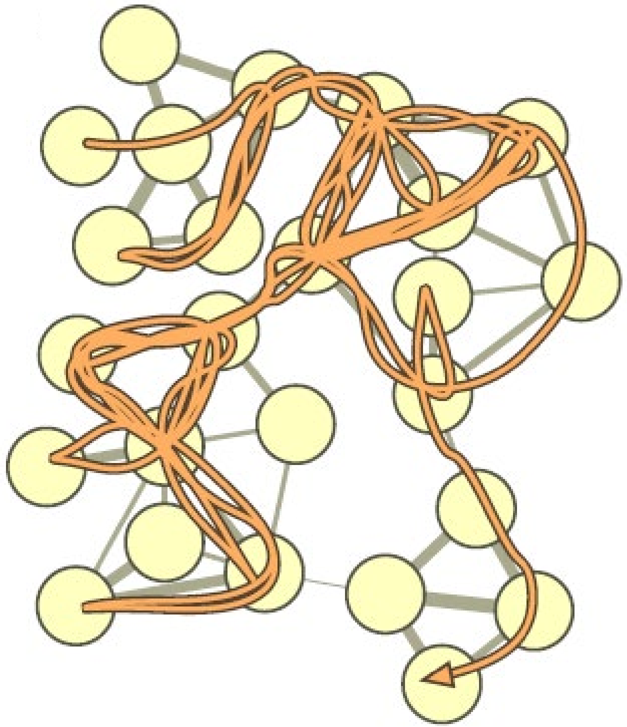
A random walker will be trapped in a community for a long time.



A number of different implementations:

- Infomap *[Rosvall and Bergstrom, 2008]*, based on information theoretic coding of random paths
- Stability of partitions *[Delvenne et al, 2010]*, based on the autocorrelation function of a signal emitted by the random walkers
- LinkRank *[Kim et al, 2010]*, extending the notion of PageRank to links
- …others…

RECAP: RANDOM WALKS ON NETWORKS



Directed, strongly connected network with N nodes, L edges, weight matrix $W = [w_{ij}]$, node out-strength $s_i^{out} = \sum_j w_{ij}$.

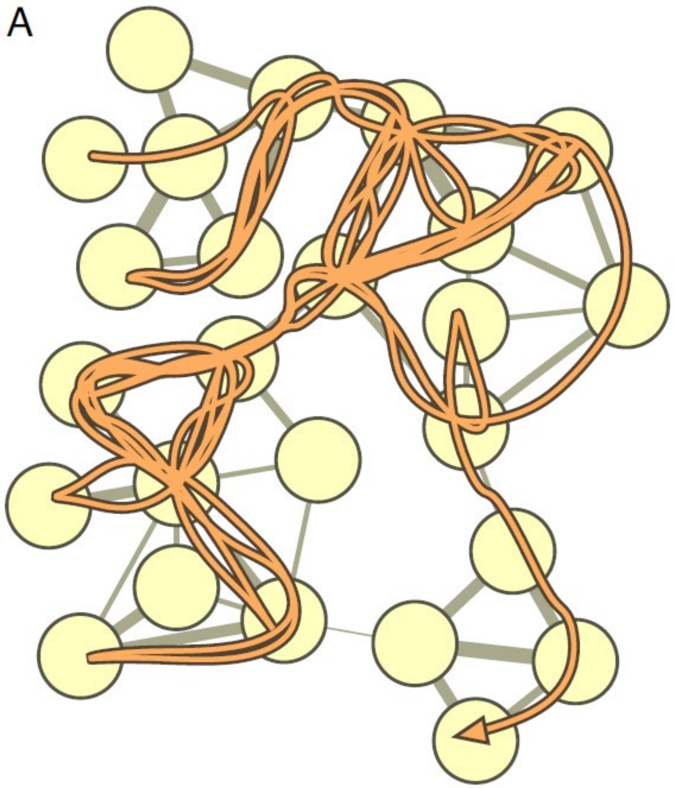
A random walker jumps from node i to j with probability

$$p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}} = \frac{w_{ij}}{s_i^{out}}$$

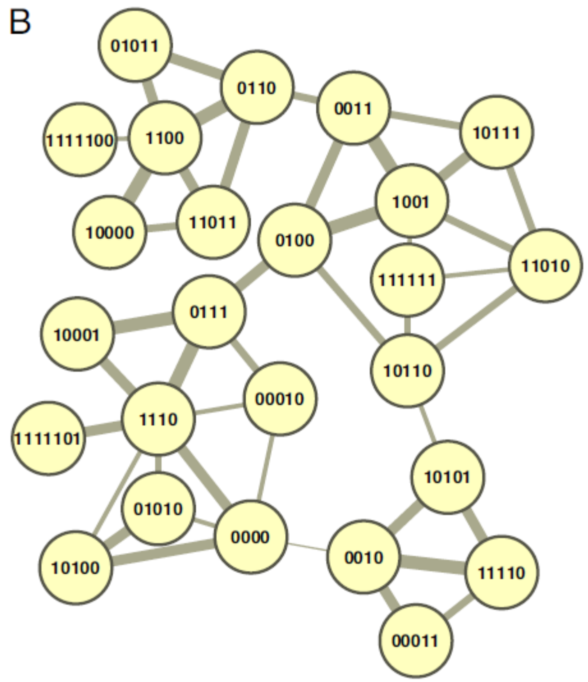
The state probability $\pi = (\pi_1 \quad \pi_2 \quad \cdots \quad \pi_N)$ evolves according to the Markov chain equation $\pi_{t+1} = \pi_t P$.

The network is strongly connected $\Rightarrow$ the transition matrix $P = [p_{ij}]$ is irreducible $\Rightarrow$ there exists a unique stationary state probability distribution $\pi = \pi P$, which is strictly positive ($\pi_i > 0$ for all i).

INFOMAP *[Rosvall and Bergstrom, 2008]*

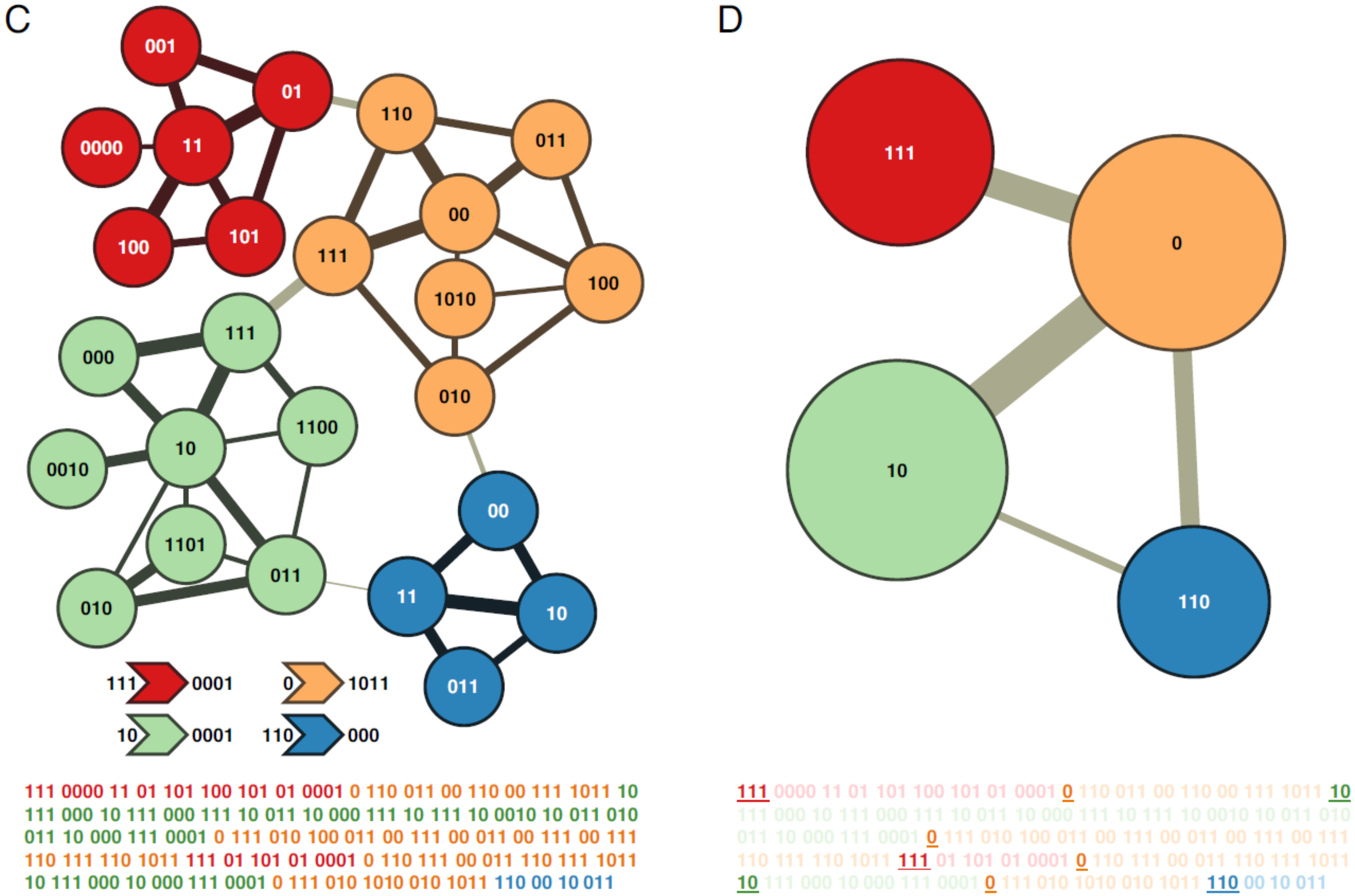


To naively describe the 71-step random walk on this 25-node network, we need 71x5=355 bits (coding each node with 5 bits).



1111100 1100 0110 11011 10000 11011 0110 0011 10111 1001 0011
1001 0100 0111 10001 1110 0111 10001 0111 1110 0000 1110 10001
0111 1110 0111 1110 1111101 1110 0000 10100 0000 1110 10001 0111
0100 10110 11010 10111 1001 0100 1001 10111 1001 0100 1001 0100
0011 0100 0011 0110 11011 0110 0011 0100 1001 10111 0011 0100
0111 10001 1110 10001 0111 0100 10110 111111 10110 10101 11110
00011

Using a Huffman code, we save space by assigning shorter codes to frequently visited nodes (=higher random walk centrality): here we only need 314 bits.



A two-level description: modules receive unique names (111,10,0,110), plus an extra code to indicate the exit (0001,0001,1011,000), and the names of nodes within clusters are reused. Here describing the 71-step walk only needs 243 bits.

The partition **M** yielding – on average – the minimal description length of a random walk is the one that minimizes this quality function (“map equation”):

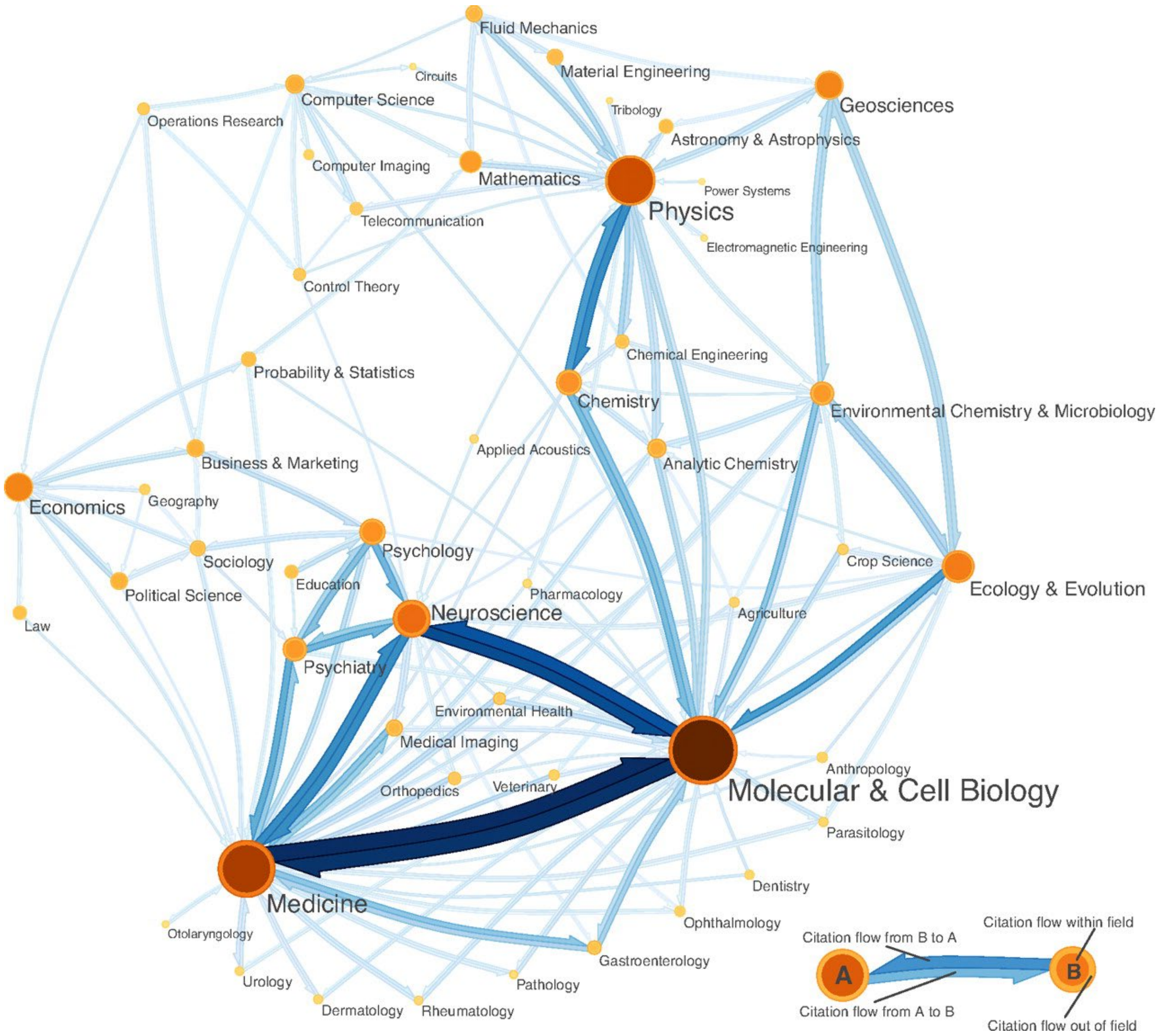
$$L(\mathbf{M}) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} p_{\circlearrowright}^{i} H(\mathcal{P}^{i}).$$

= avg bits per step for describing (inter-community + intra-community) dynamics

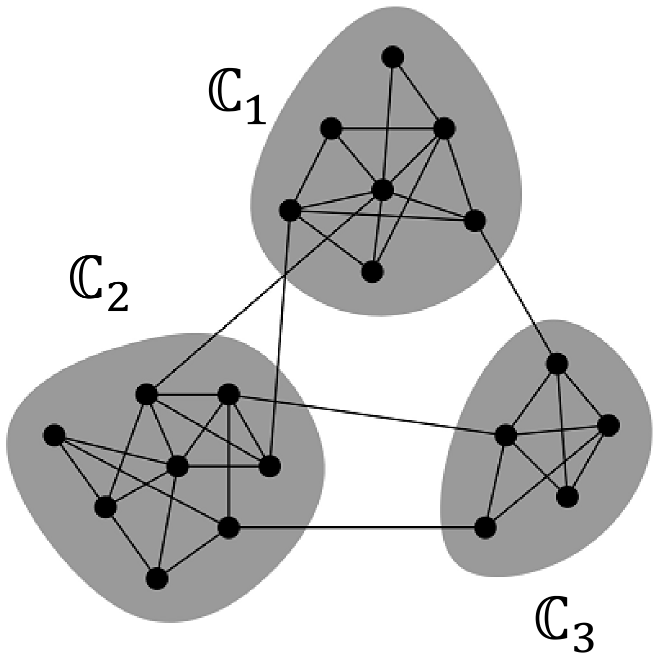
The partition attaining *min* $L(\mathbf{M})$ is taken as the “best” partition, as small $L(\mathbf{M})$ implies long persistence within modules.

(implementations with complexity $O(NlogN)$, with strategies similar to Louvain)

Applying Infomap to a citation network (6,128 journals, 6M+ citations) reveals 88 thematic modules:



NETWORK + PARTITION = *LUMPED* MARKOV CHAIN



$\mathbb{C}_c$ is a set of nodes (a "candidate" community), and $\mathbb{P}_q = \{\mathbb{C}_1, \mathbb{C}_2, \dots, \mathbb{C}_q\}$ is a partition.

$\mathbb{P}_q$ is coded by the $N \times q$ binary collecting matrix $H = [h_{ic}]$:

$$h_{ic} = 1 \quad \Leftrightarrow \quad i \in \mathbb{C}_c$$

The dynamics of the random walker at this aggregate scale ("meta-network") is described, at stationarity ($\pi_0 = \pi$), by the q-state lumped Markov chain

$$\Pi_{t+1} = \Pi_t U \quad \text{where} \quad U = [\mathrm{diag}(\pi H)]^{-1} H' \mathrm{diag}(\pi) P H$$

u_{cd} = probability that the random walker is at time $t+1$ in any of the nodes of $\mathbb{C}_d$ provided it is in t in any of the nodes of $\mathbb{C}_c$

PERSISTENCE PROBABILITIES *[Piccardi, PLoS ONE, 2011]*

The diagonal terms u_{cc}, $i = 1,2,\dots,q$, of the lumped Markov matrix U are called PERSISTENCE PROBABILITIES.

Significant communities are expected to have large persistence probability u_{cc} (thus large escape time $\tau_c = (1 - u_{cc})^{-1}$).

$$u_{cc} = \frac{\sum_{i,j \in \mathbb{C}_c} \pi_i p_{ij}}{\sum_{i \in \mathbb{C}_c} \pi_i} = \mathit{fraction\ of\ time\ spent\ by\ the\ random\ walker\ on\ the}\ \frac{\mathit{links}}{\mathit{nodes}}\ \mathit{of\ community}\ \mathbb{C}_c$$

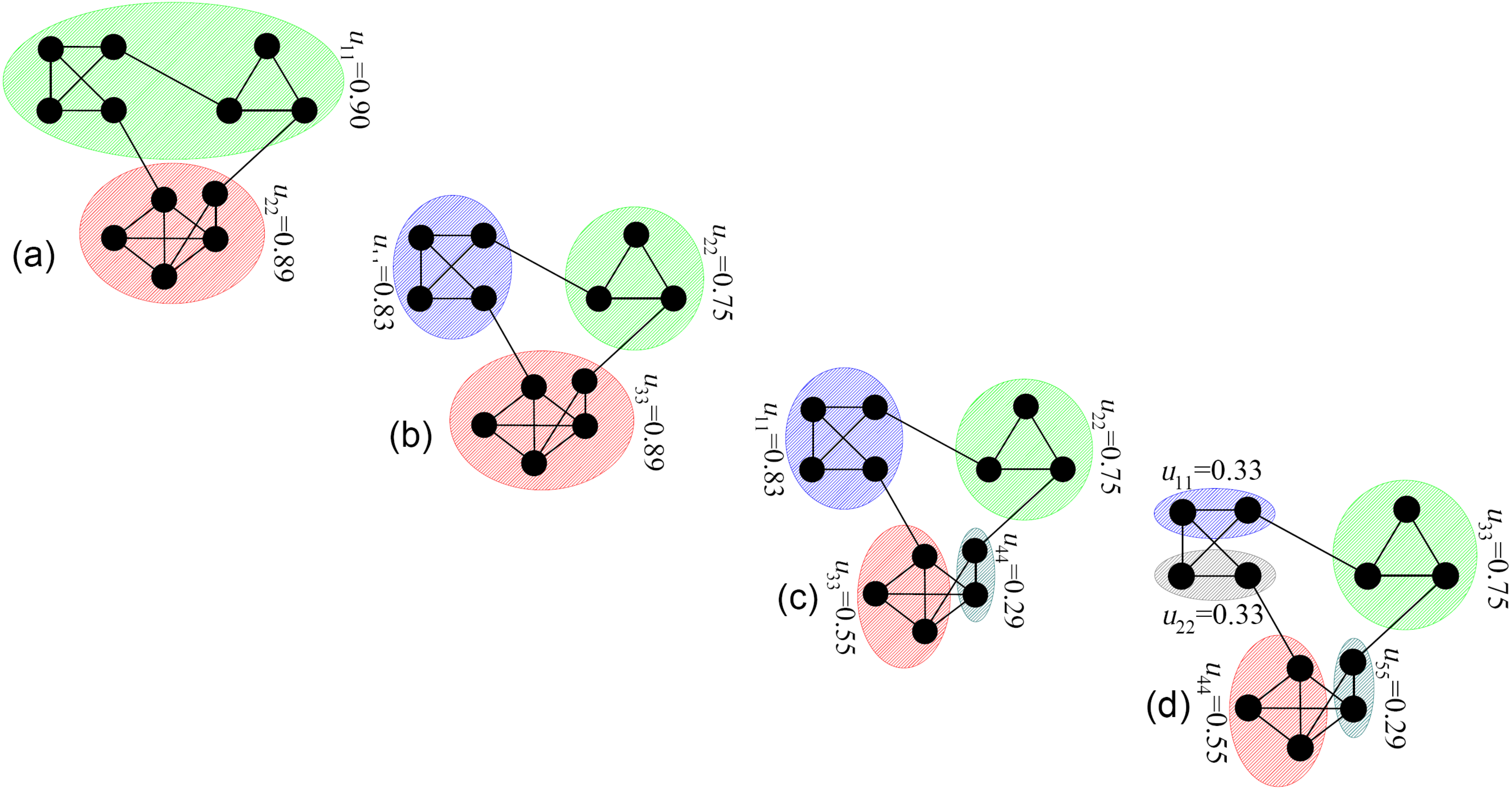
If the network is undirected:

$$u_{cc} = \frac{\sum_{i \in \mathbb{C}_c} s_i^{int}}{\sum_{i \in \mathbb{C}_c} s_i} = \frac{\mathit{total\ internal\ strength}}{\mathit{total\ strength}}\ \mathit{of\ community}\ \mathbb{C}_c = \mathit{fraction\ of\ strength\ internally\ directed}$$

If the network is undirected and unweighed:

$$u_{cc} = \frac{\mathit{total\ internal\ degree}}{\mathit{total\ degree}}\ \mathit{of\ community}\ \mathbb{C}_c > 0.5 \Longleftrightarrow \mathbb{C}_c$$ is a "community" according to Radicchi et al. [PNAS, 2004]

Persistence probabilities reveal significant communities / partitions.



α-COMMUNITIES AND α-PARTITIONS

Set a quality level $0 < \alpha < 1$.

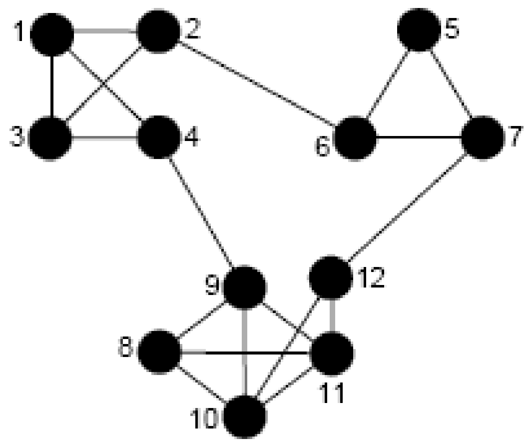
- $\mathbb{C}_c$ is an α-community if the persistence probability $u_{cc} \geq \alpha$.
- $\mathbb{P}_q = \{\mathbb{C}_1, \mathbb{C}_2, \dots, \mathbb{C}_q\}$ is an α-partition if $\mathbb{C}_1, \mathbb{C}_2, \dots, \mathbb{C}_q$ are α-communities (i.e., $\min_c u_{cc} \geq \alpha$).

A strategy for community analysis:

- set the quality level α
- generate a set of "good" candidate partitions, with different number q of clusters (many algorithms are available)
- take the α-partition with the largest q (i.e., the finest decomposition with the desired quality level)

Remark: the "quality" (significance) of each individual community is simultaneously assessed.

Finding communities: the "persistence probabilities' diagram (PPD)"

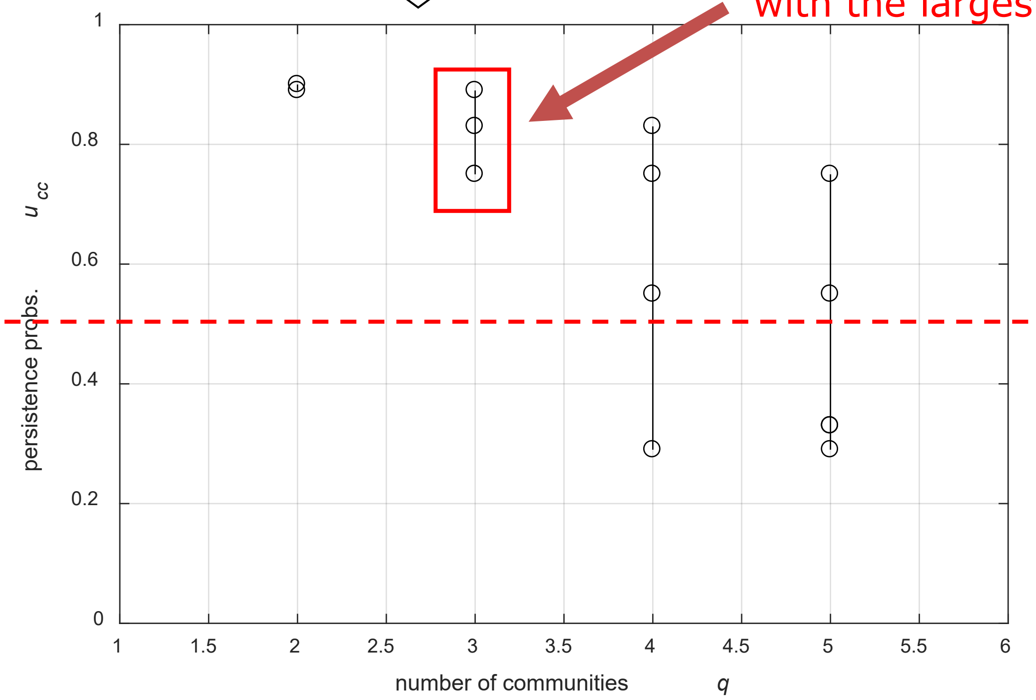


any "partitions generator":

$\mathbb{P}_2, \mathbb{P}_3, \mathbb{P}_4, \dots$

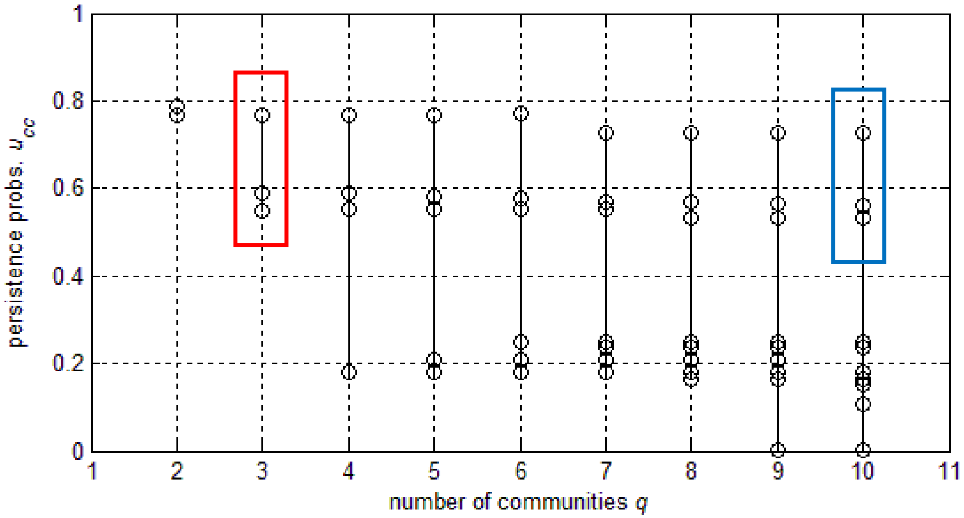
2: This is the α-partition with the largest q

1: We set a quality level, e.g. $\alpha = 0.5$



Example: Communities in the World Trade Network (WTN, 2008)

The network can only be decomposed into 3 significant clusters, if a reasonably high quality level is sought for (e.g., $\alpha = 0.5$).



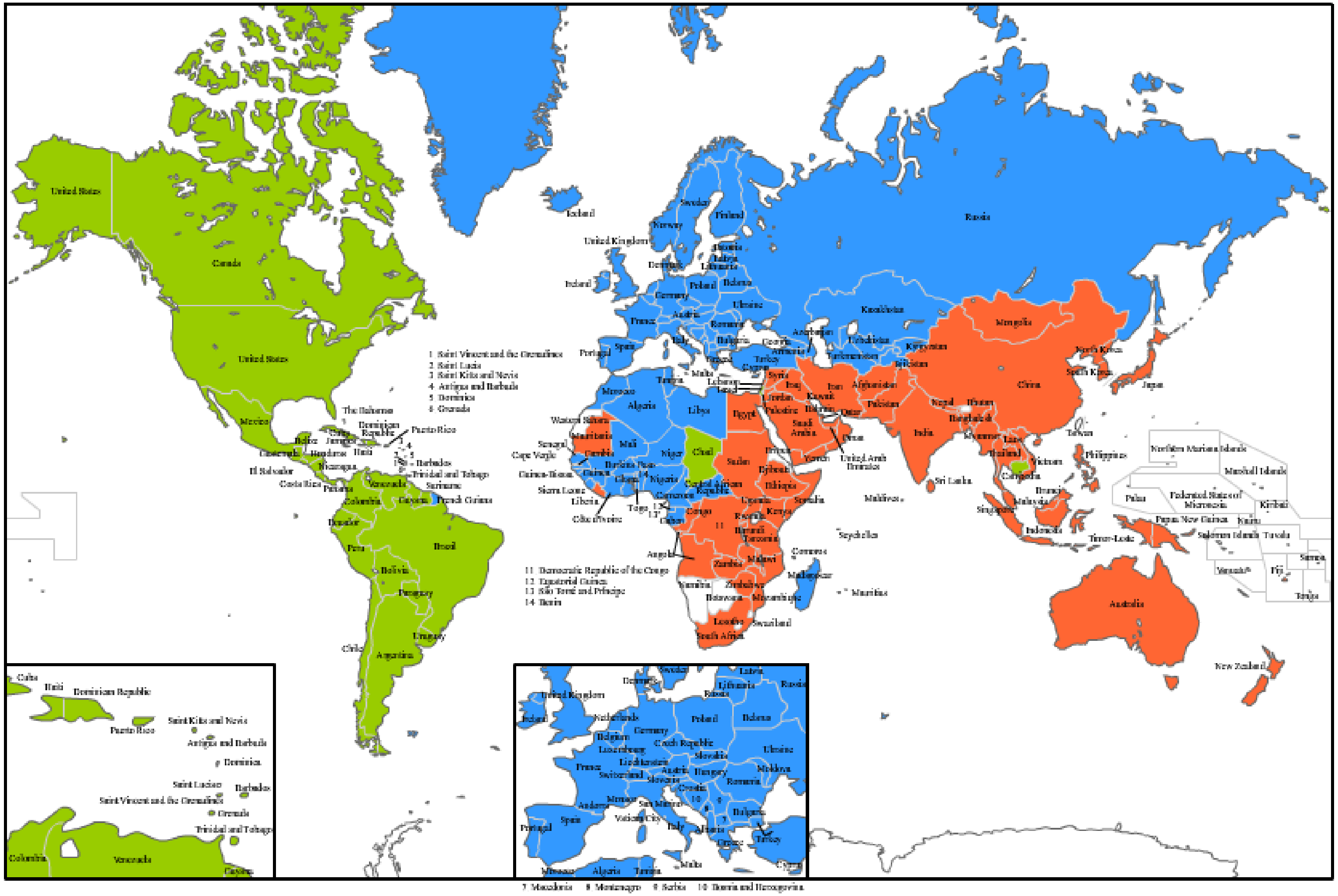
$\mathbb{C}_1$ ($u_{11} = 0.77$):
Europe (incl. ex-USSR countries) + half of Africa

$\mathbb{C}_2$ ($u_{22} = 0.59$):
Asia + Australia + half of Africa

$\mathbb{C}_3$ ($u_{33} = 0.55$):
Americas

- No further decomposition is significant $\Rightarrow$ the partition is more or less "trivial"
- The 3 clusters have a "stable core": taking a larger q simply "peels off" a few peripheral countries
- "Europe" ($\mathbb{C}_1$) is the only community with large persistence probability $\Rightarrow$ even the 3-way partition is "weak" (i.e., the network is weakly clusterized)

$\mathbb{C}_1$ $(u_{11} = 0.77)$: Europe (incl. ex-USSR countries) + half of Africa
$\mathbb{C}_2$ $(u_{22} = 0.59)$: Asia + Australia + half of Africa
$\mathbb{C}_3$ $(u_{33} = 0.55)$: Americas



Example: Communities in criminal networks: the Infinito case

[Calderoni, Brunetto & Piccardi, Social Networks, 2017]

- "Operazione Infinito" (2011): large law enforcement operation (more than 150 people arrested)

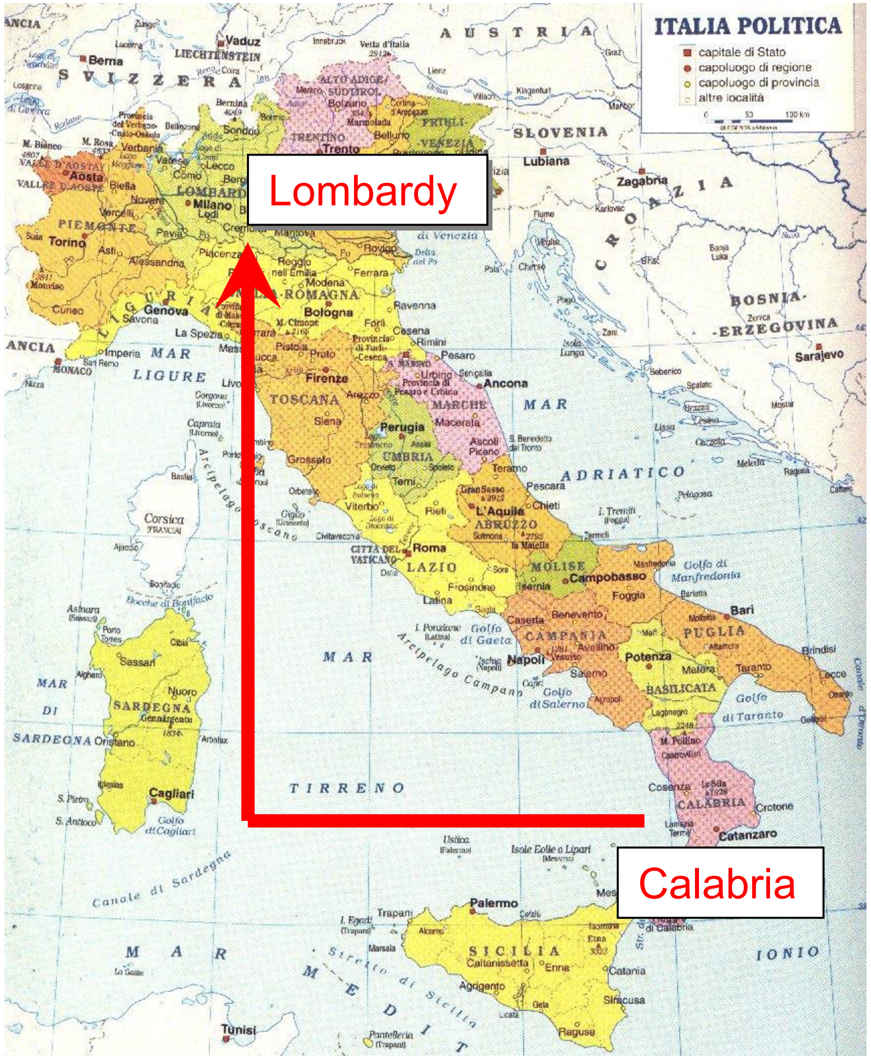
- establishment of several 'Ndrangheta groups in Lombardy

- structure of the criminal organization: formal membership to a *Locale*

- from the investigations: records meetings/participants

Aims of network analysis:

- understanding the structure *(is the organization really clustered?)*
- helping future investigations *(could the* Locali *membership be predicted?)*



The Infinito network

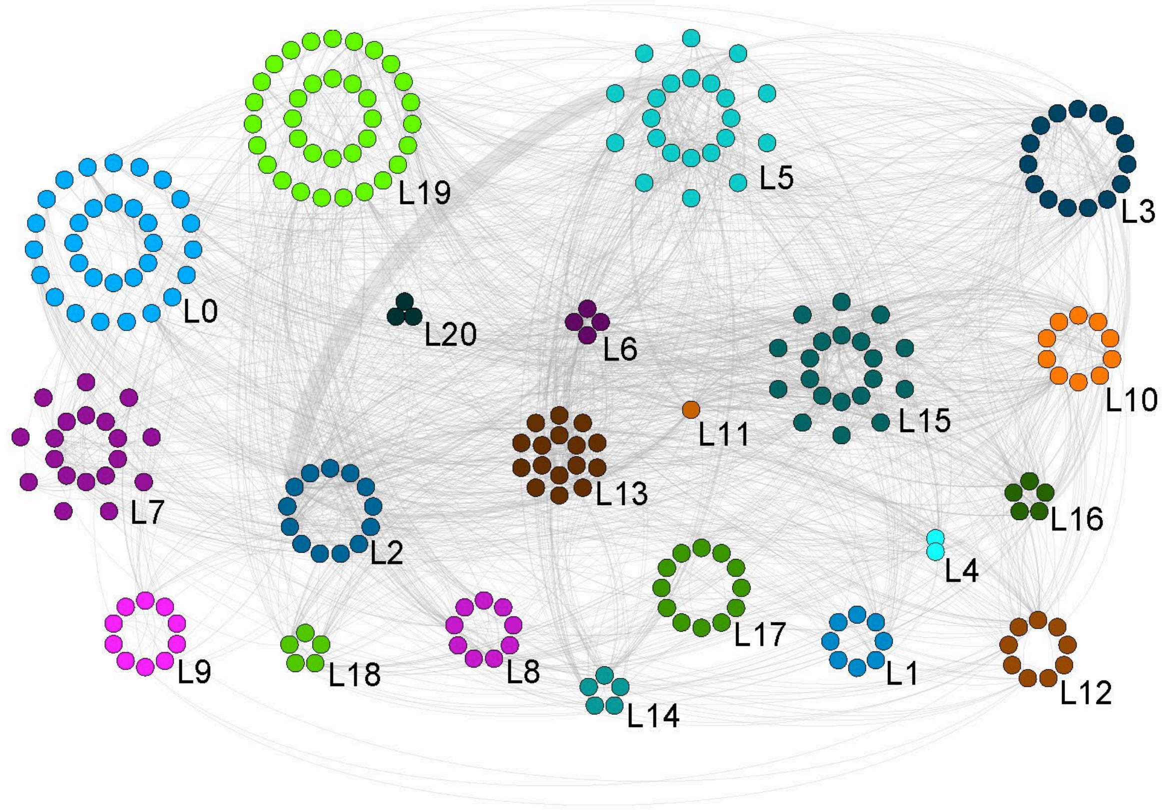
records meetings/participants

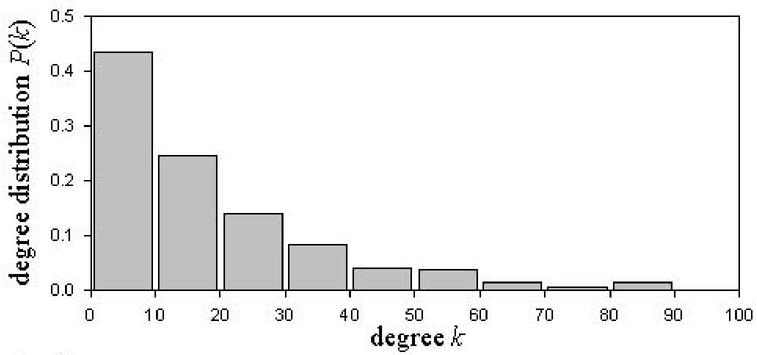
⟹

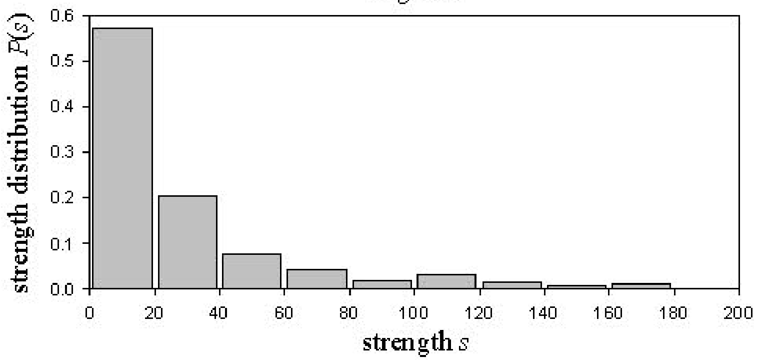
bipartite (two-mode) network

⟹

projection onto the set of participants







$N = 254$ nodes; $L = 2132$ links; density $\rho = 0.066$

undirected, weighted network

(w_{ij}=n. of co-participations in meetings)

Testing the significance of the *Locali* partition

"Operazione Infinito": 177 individuals (70%) are associated to 17 *Locali* in Lombardy

Are the *Locali* significant as communities (i.e., cohesive)?

C_k = subgraph induced by *Locale* k (with N_k nodes)

We quantify the cohesiveness of C_k by:

α_k = persistence probability of C_k =
prob. that a random walker in C_k remains in C_k at the next step

In undirected networks: α_k = fraction of the strength of the nodes of C_k directed within C_k

$$\alpha_k = \frac{\sum_{i \in C_k} \sum_{j \in C_k} w_{ij}}{\sum_{i \in C_k} \sum_{j \in \{1,2,\dots,N\}} w_{ij}}$$

The larger α_k, the larger the cohesiveness of C_k $\Longrightarrow$ threshold $\alpha_k > 0.5$

TABLE I. TESTING THE *Locali* PARTITION

	locale	N_k	α_k	z_k
L0	*Not specified*	*31*	*0.08*	*-3.15*
L1	*Not affiliated*	*8*	*0.03*	*-0.84*
L2	Bollate	13	0.25	1.31
L3	Bresso	15	0.39	2.72
L4	Canzo	2	0.10	0.47
L5	Cormano	22	0.41	3.96
L6	Corsico	4	0.12	0.21
L7	**Desio**	**19**	**0.63**	**6.40**
L8	Erba	9	0.27	2.44
L9	**Giussano**	**10**	**0.63**	**5.26**
L10	Legnano	10	0.20	0.77
L11	Limbiate	1	0	
L12	Mariano Comense	9	0.27	1.40
L13	**Milano**	**16**	**0.62**	**5.78**
L14	Pavia	5	0.13	0.25
L15	Pioltello	20	0.43	3.83
L16	Rho	5	0.18	0.78
L17	**Seregno**	**12**	**0.93**	**8.73**
L18	Solaro	5	0.06	-0.42
L19	*Calabria locali*	*35*	*0.19*	*-0.97*
L20	*Brescia*	*3*	*0.17*	*0.98*

Remark: in all nets α_k tends to increase (from 0 to 1) as N_k grows $\Longrightarrow$ need to check for statistical significance:

$$z_k = \frac{\alpha_k - \mu(\bar{\alpha}_k)}{\sigma(\bar{\alpha}_k)}$$

$\mu(\bar{\alpha}_k)$, $\sigma(\bar{\alpha}_k)$: mean & st. dev. of the persistence probabilities of all (connected) subnets of size N_k

Only 4 *Locali* (over 17) are cohesive as communities ($\alpha_k > 0.5$, with $z_k > 3$).

Overall, no evidence of strong clusterization *based on the Locali*.

Community analysis: max-modularity

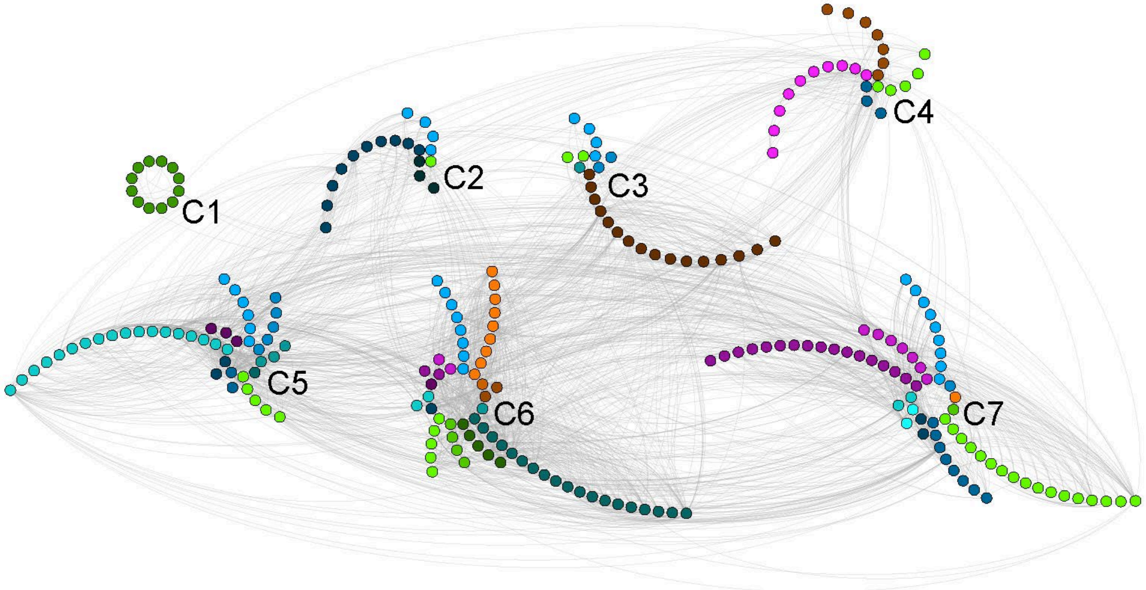
$$\max\ Q = \frac{1}{2s}\sum_{k=1,2,\ldots K}\sum_{i,j\in C_k}\left(w_{ij} - \frac{s_i s_j}{2s}\right) \Longrightarrow \text{partition } C_1, C_2, \ldots, C_K$$

RESULTS OF MAX-MODULARITY COMMUNITY ANALYSIS

	N_k	α_k	z_k
C1	12	0.93	9.07
C2	18	0.72	7.79
C3	25	0.66	9.85
C4	25	0.63	9.11
C5	45	0.68	8.20
C6	62	0.78	8.30
C7	67	0.67	5.72

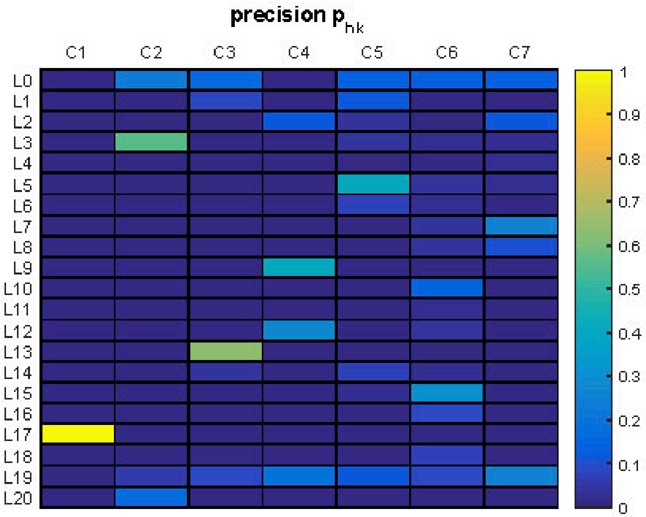
- $Q_{max} = 0.48$
- $K = 7$ communities
- all (very) cohesive ($\alpha_k > 0.5$, with $z_k > 3$)

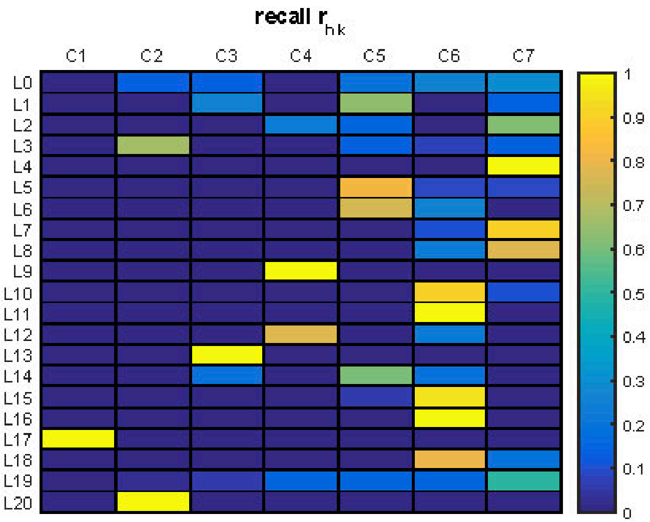
How do communities $\{C_1, C_2, \ldots, C_7\}$ relate to the *Locali* $\{L_2, L_3, \ldots, L_{18}\}$?



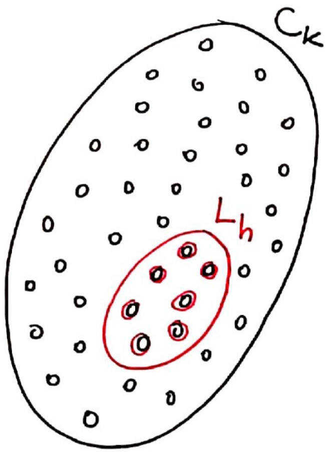
Precision/Recall analysis ("set matching")

$$p_{hk} = \frac{m_{hk}}{|C_k|}, \quad r_{hk} = \frac{m_{hk}}{|L_h|}, \quad m_{hk}\text{=n. nodes in locale } L_h \text{ and in community } C_k$$



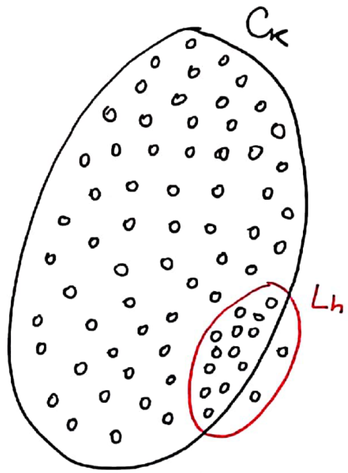


Single *locali* weekly matches with communities – but we note that...



$$r_{hk} = \frac{m_{hk}}{|L_h|} = 1$$

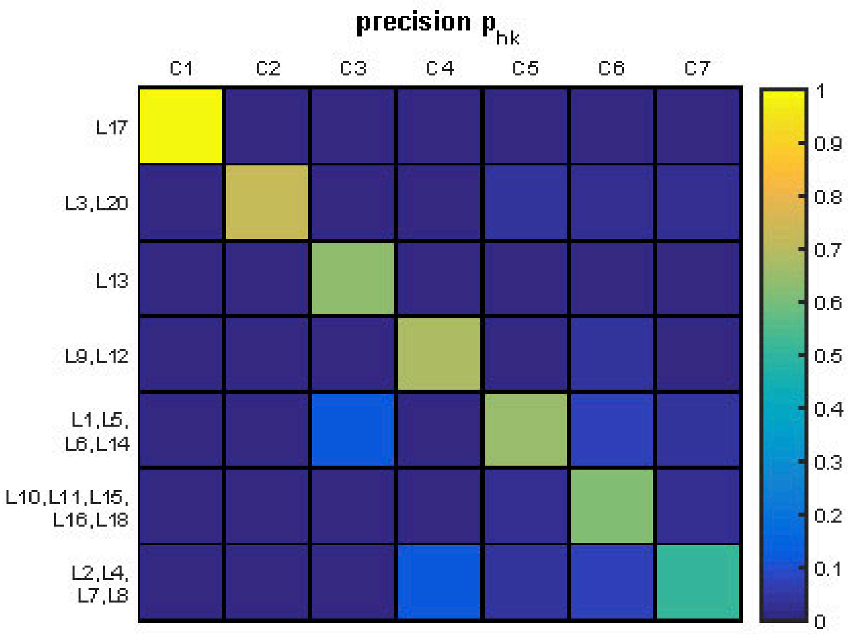
$$p_{hk} = \frac{m_{hk}}{|C_k|} \; small$$

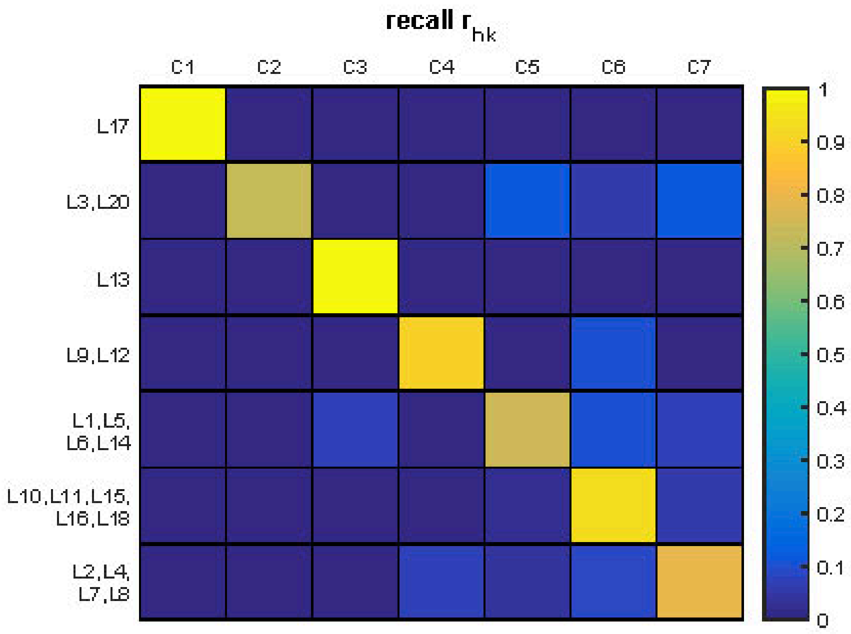


$$r_{hk} = \frac{m_{hk}}{|L_h|} \cong 1$$

$$p_{hk} = \frac{m_{hk}}{|C_k|} \; small$$

If we suitably aggregate *locali*…





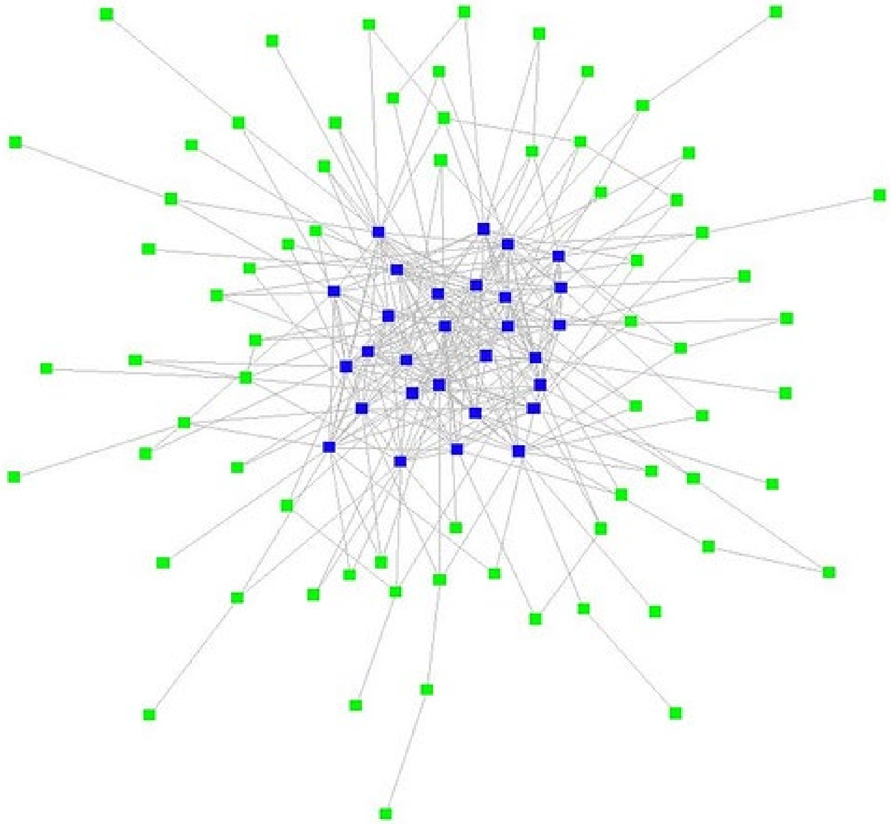
...we discover that, given the strong clusterization, communities are in fact single *locali* or mostly unions of *locali*.

Core-Periphery Analysis

CORE-PERIPHERY ANALYSIS

Core-periphery paradigm: the network is the union of a dense core with a sparsely connected periphery.

Origin in the 70's in economics (unequal economic growth/development of countries) and social sciences (elites and power), recent applications in communication networks, biology, etc.



Core-periphery analysis:

- Assess whether the network does have a core-periphery structure (i.e., is there a central core through which most of the network flow passes?).
- Assign each node to the relevant subnetwork.

Connections with centrality measures, but main focus on the whole network structure.

Block-modelling *[Borgatti & Everett, Soc. Networks, 1999]*

The ideal core-periphery network structure: *"...core nodes are adjacent to other core nodes, core nodes are adjacent to some periphery nodes, periphery nodes do not connect to other periphery nodes..."*

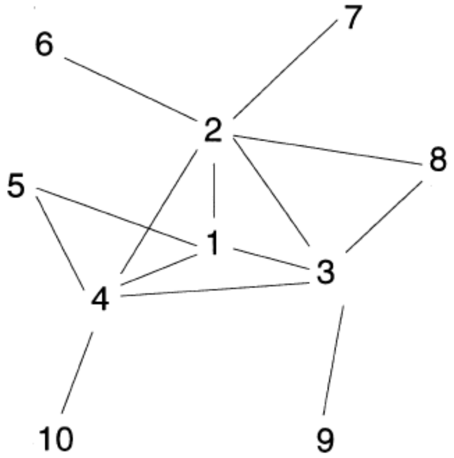


Fig. 1. A network with a core/periphery structure.

complete (all-to-all)

	1	2	3	4	5	6	7	8	9	10
1		1	1	1	1	0	0	0	0	0
2	1		1	1	0	1	1	1	0	0
3	1	1		1	0	0	0	1	1	0
4	1	1	1		1	0	0	0	0	1
5	1	0	0	1		0	0	0	0	0
6	0	1	0	0	0		0	0	0	0
7	0	1	0	0	0	0		0	0	0
8	0	1	1	0	0	0	0		0	0
9	0	0	1	0	0	0	0	0		0
10	0	0	0	1	0	0	0	0	0	

fully disconnected

Fitting the ideal structure to our concrete network:

- find the 2-way partition that maximizes 1's among core nodes and 0's among periphery nodes (can be cast as an optimization problem).

Drawbacks: unknown significance of the obtained partition; too crude separation.

k-core (k-shell) decomposition

The k-core is the (maximal) subgraph S whose nodes have (internal) $deg_S \geq \mathrm{k}$.

The k-shell is the set of nodes belonging to the k-core but not to the (k+1)-core.

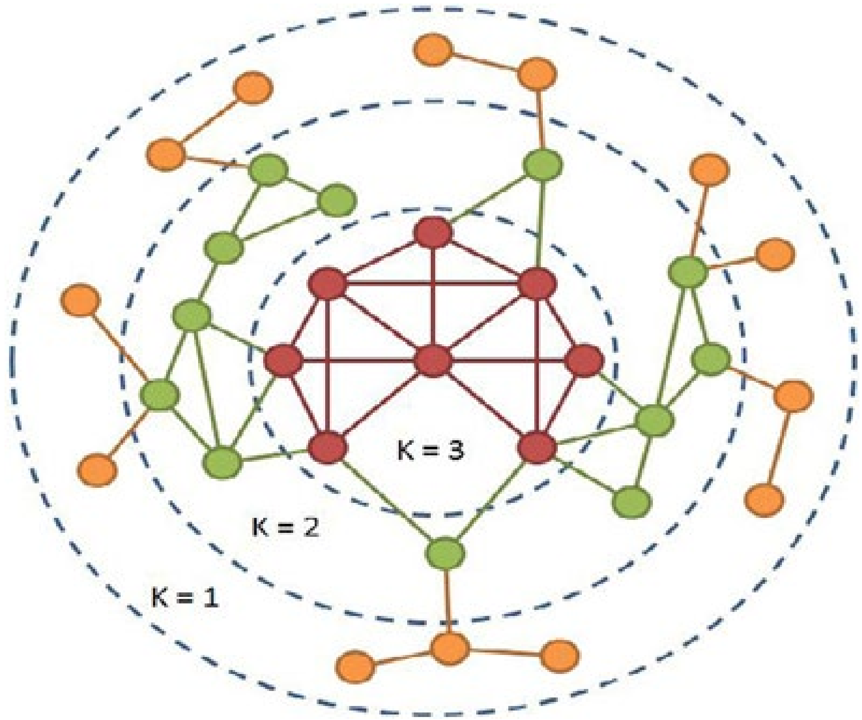
Thus the network is organized into "concentric" layers, the k-shells. The union of all k'-shells with $k' \geq k$ is the k-core.

Decomposition algorithm:

- put in the 1-shell - and remove - the degree-1 nodes, as well as, recursively, those having degree 1 after removal of the former;

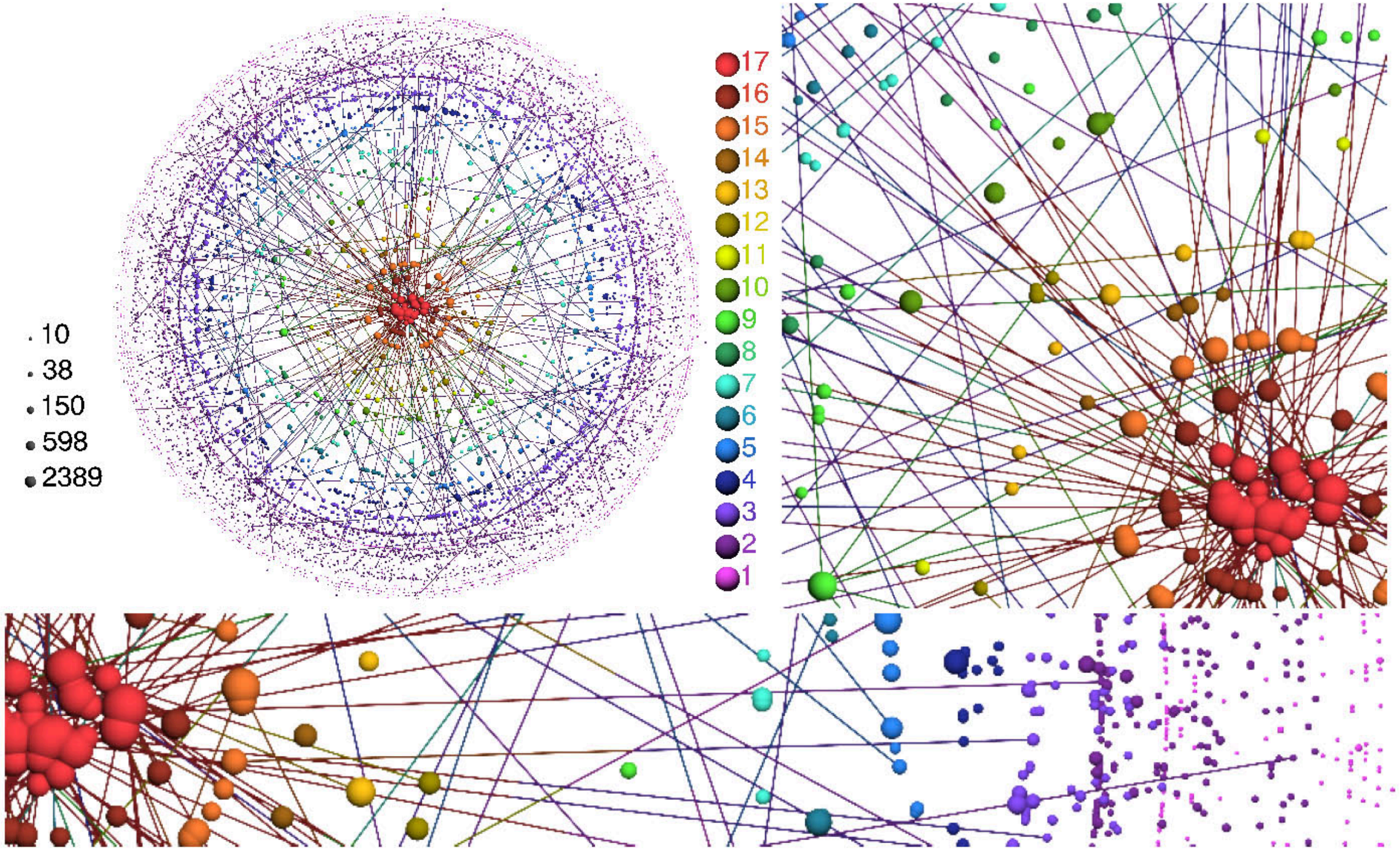
- put in the 2-shell - and remove - the degree-2 nodes, as well as, recursively, those having degree ≤2 after removal of the former;

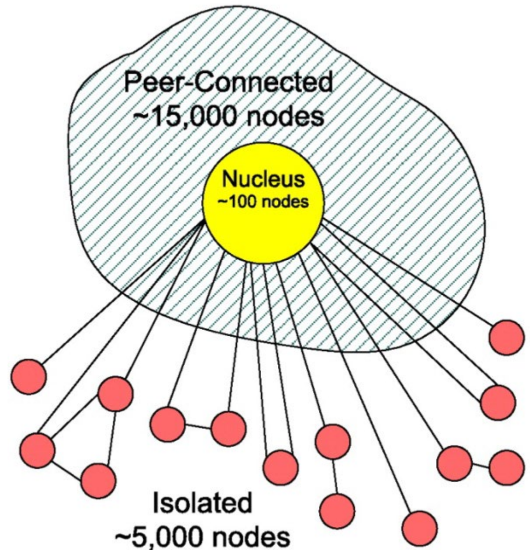
- etc...



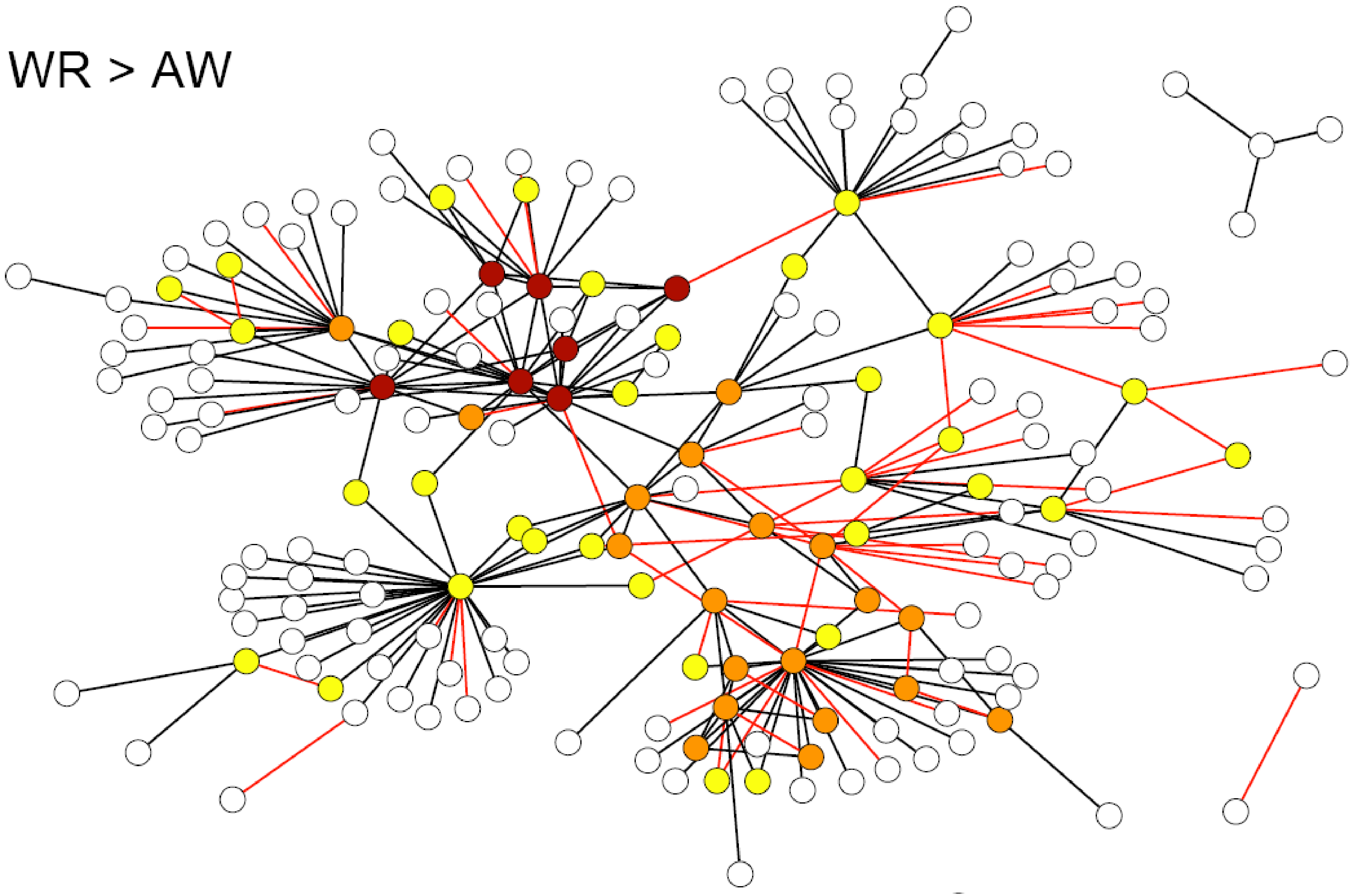
The k-coreness of a node (=the k-shell it belongs to) is a measure of centrality.

Example: k-core decomposition of the Internet (autonomous system level)





Example: k-core decomposition of a criminal network (mafia groups in Northern Italy)

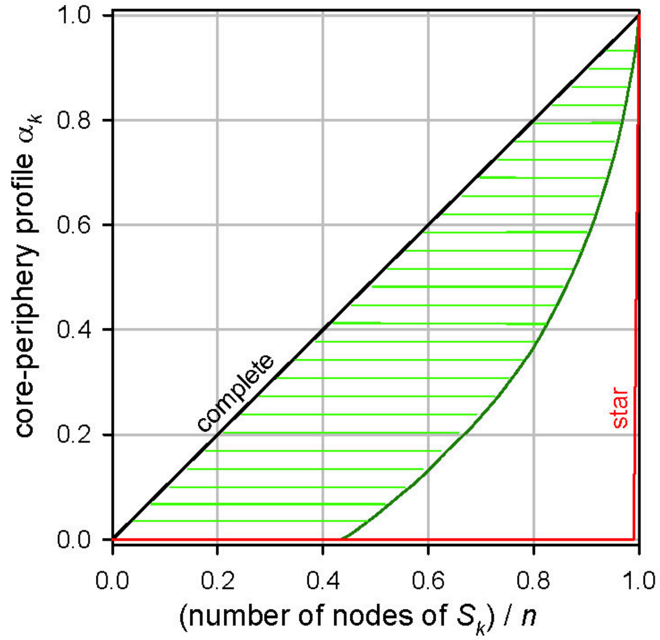


Core-Periphery profile

[Della Rossa, Dercole, Piccardi, Scientific Rep., 2013]

A heuristic procedure for ordering the nodes from to the periphery to the core:

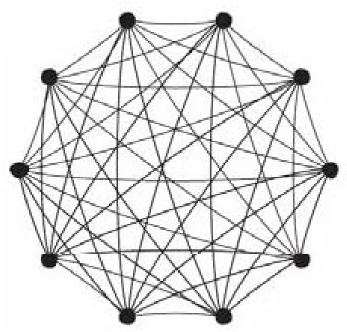
- start by the node i with minimal strength
- generate a sequence of sets $\{i\} = S_1 \subset S_2 \subset \cdots \subset S_N = \{1,2,\dots,N\}$ by adding, at each step, the node attaining the minimal persistence probability $\alpha_1, \alpha_2, \dots, \alpha_N$.



The sequence $0 = \alpha_1 \le \alpha_2 \le \cdots \le \alpha_N = 1$ is the Core-Periphery profile (and α_k is the coreness of the node inserted at step k).

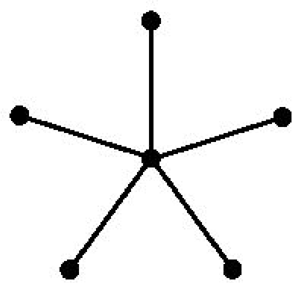
The Core-Periphery score C is the ([0,1]-normalized) area between the Core-Periphery profile and the profile of the complete network.

complete nets



$C = 0$

star nets



$C = 1$

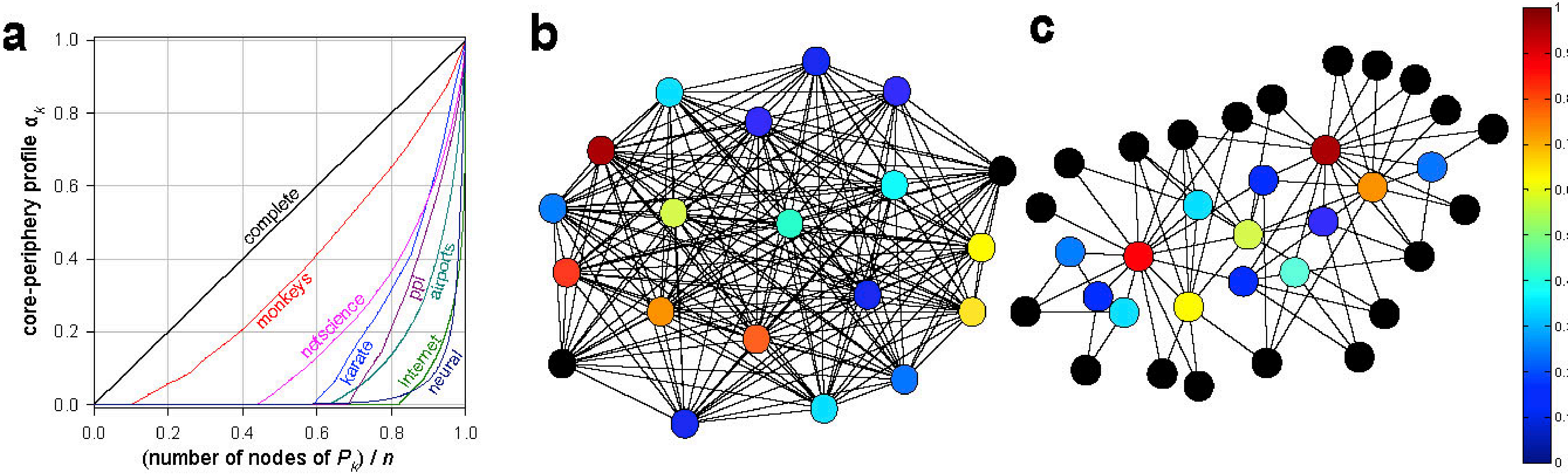
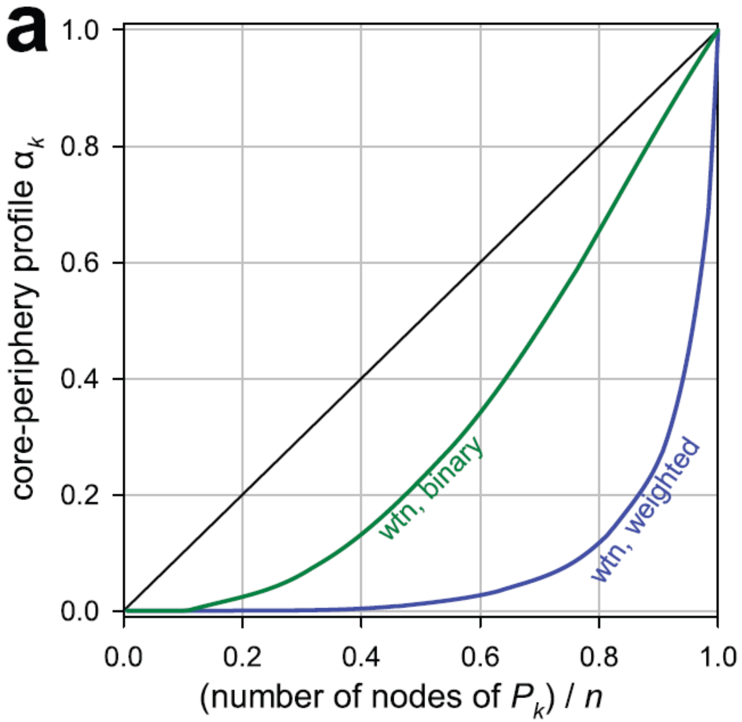


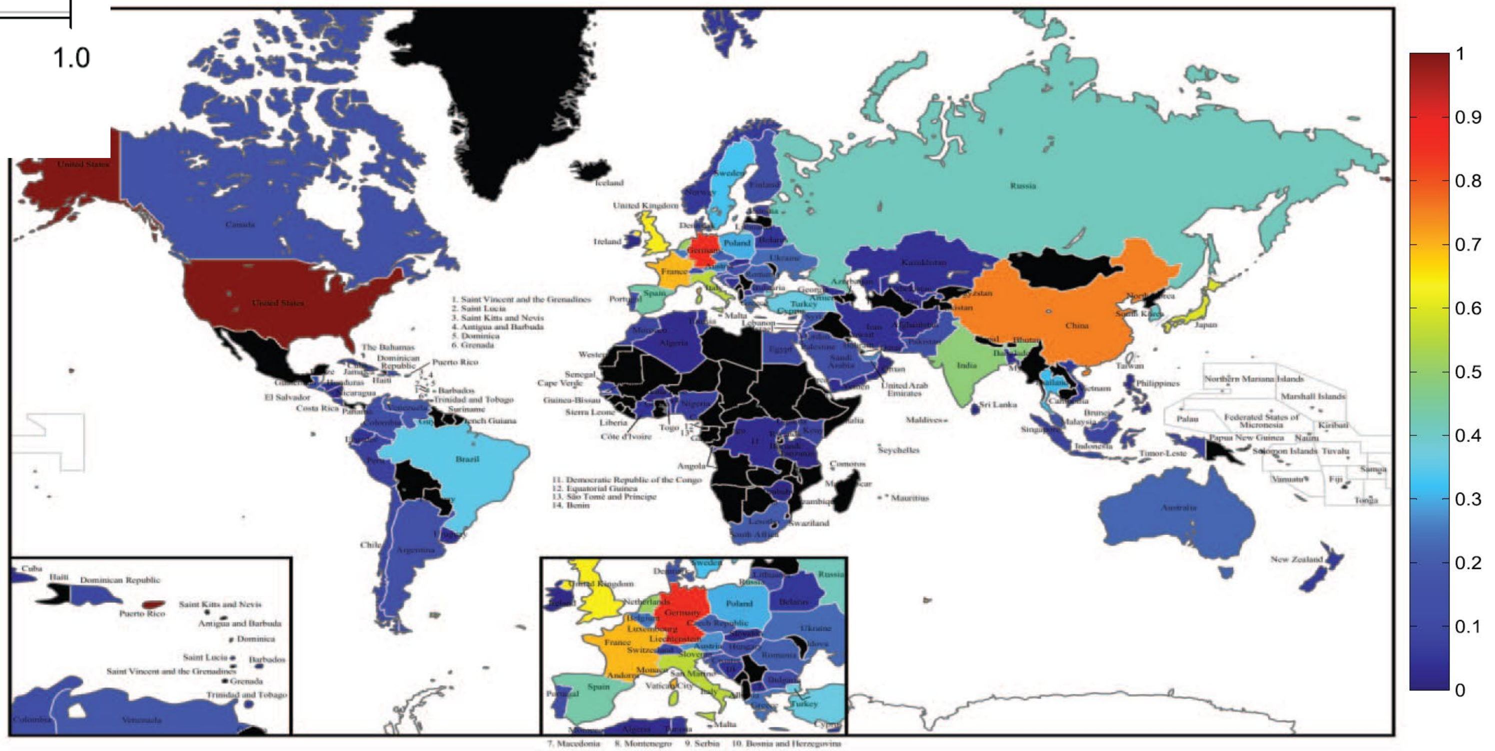
Figure 2 | **Core-periphery analysis of real-world networks.** (a). The core-periphery profiles of the networks describing: the social interactions within a troop of *monkeys*, $n = 20$ (graph in panel (b)); the friendship among the members of a *karate* club, $n = 34$ (graph in panel (c)); the coauthorships among scientists working on networks (*netscience*), $n = 379$; the protein-protein interaction (*ppi*) network of *Saccharomyces cerevisiae*, $n = 1458$; the international *airports* network, $n = 2868$; the *Internet* in 2006, at the level of autonomous systems, $n = 11745$; and the *neural* network of the worm *Caenorhabditis elegans*, $n = 239$. In graphs (b) and (c), nodes are coloured according to their coreness: p-nodes ($\alpha_k = 0$) are in black.



The World Trade Network:

- is complete-like, if weights are neglected (binary - topology only)

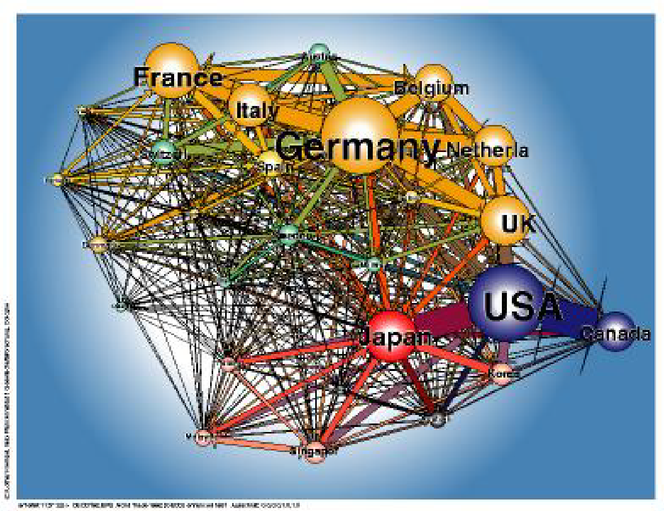
- is star-like, if weights are accounted for (only United States, Germany, China, France, United Kingdom, Japan, Italy, and the Netherlands, in order, have coreness $\alpha_k > 0.5$).



Example: Complexity, Centralization, and Fragility in Economic Networks

[Piccardi and Tajoli, 2018]

How fragile is the world economy?



world trade network, 1992 (www.mpi-fg-koeln.mpg.de)

Given the increasing globalization of economic systems, will economic shocks have widespread diffusion to all countries?

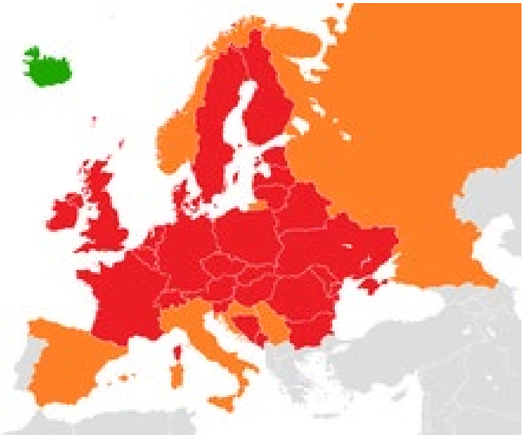
Two contrasting effects of the increased number of economic links:

- Diversification, averaging effects, more resilience
- More connections, more effective shock propagation

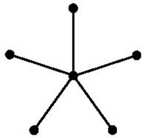
The international financial crisis (2007-2008) and the European debt crisis (2009-2010) suggest that indeed most of the world countries are highly exposed.

Broad (economic) impact of localized (non-economic) events:

Eyjafjallajökull eruption (2010)

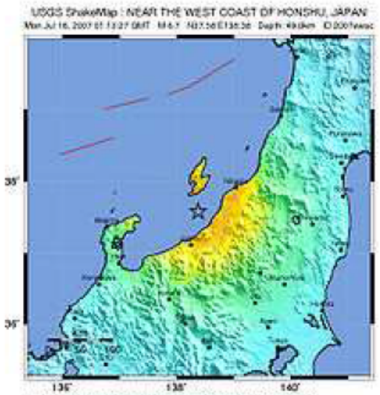


impact on flight traffic



Japan earthquake (2007)

We analyzed the effect of business network damaged by an earthquake of July 16th, 2007.



Magnitude 6.6, not very big .

But it has huge impact on the economy.

All Japanese car makers (Toyota, Honda, Nissan, Matsuda, Mitsubishi...) stopped their production lines for about a week.

It causes a loss of 130,000 car production (=3 billion US dollars)

This was due to a breakdown of a small factory called "Riken" producing "piston rings"

Riken is the leading company

50% share in Japan

20% in the world

Very high-tech

Riken is an indispensable company. No alternative company to make such high quality piston rings for luxury cars.

No body paid attention to such a local company before the earthquake

Misako Takayasu,
plenary talk @ CCS 2016, Amsterdam

We focus on product trade networks and explore the relationship between product complexity and network centralization.

Research question:

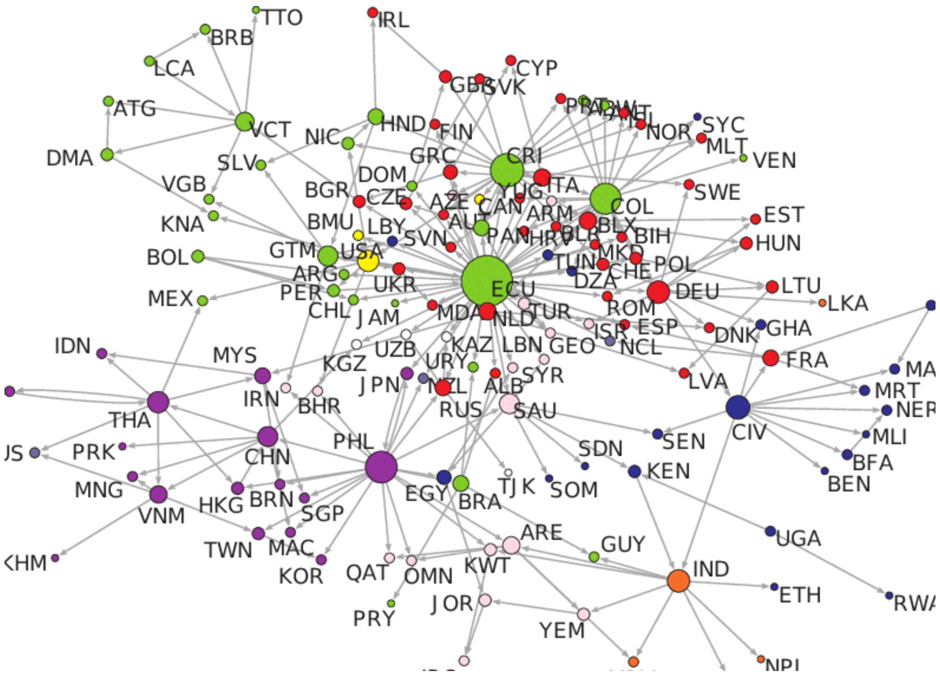
Are complex (high-tech) products distributed through centralized (hence more fragile) networks?

DATA

- Inter-country trade (year 2014) among 223 countries (CEPII-BACI database).
- HS 4-digit classification (1,242 products, partitioned into 15 Sections).

Examples:

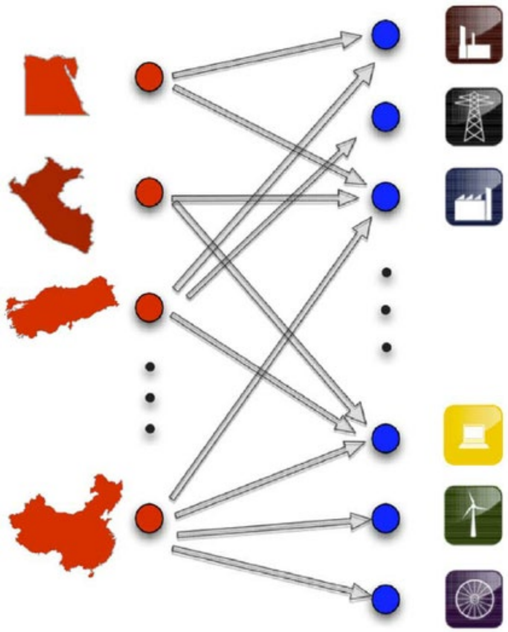
- Products #1211: "Plants and parts of plants, including seeds and fruits" (Section "Vegetable Products")
- Product #8513: ”Portable electric lamps designed to function by their own source of energy” (Section ”Machinery/Electrical”)



example: trade network of bananas

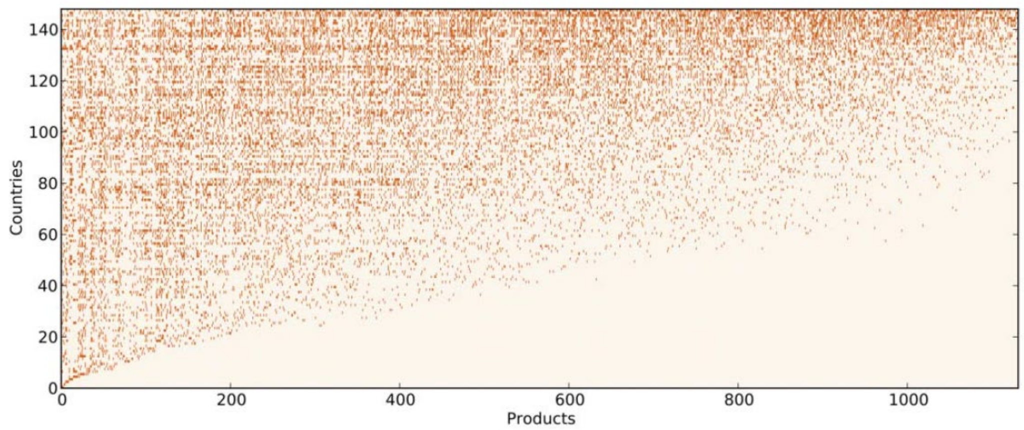
De Benedictis et al., Global Econ J, 2014

Aggregating data: the Country-Product bipartite ("two-mode") network



Caldarelli et al., Plos One, 2012

$E = [e_{cp}]$: trade matrix,
export (USD) of product p from country c



$M = [m_{cp}]$: binarized trade matrix,
$m_{cp} = 1$ if $r_{cp} > 1$
(Revealed Comparative Advantage)

MEASURING PRODUCT COMPLEXITY

"Traditional" measures:

- Technological class: products are partitioned into 5 categories (qualitative, based on expertize – Lall, 2000):

  PP: primary product
  RB: resource-based manufacture
  LT: low-technology manufacture
  MT: medium-technology manufacture
  HT: high-technology manufacture

  complexity

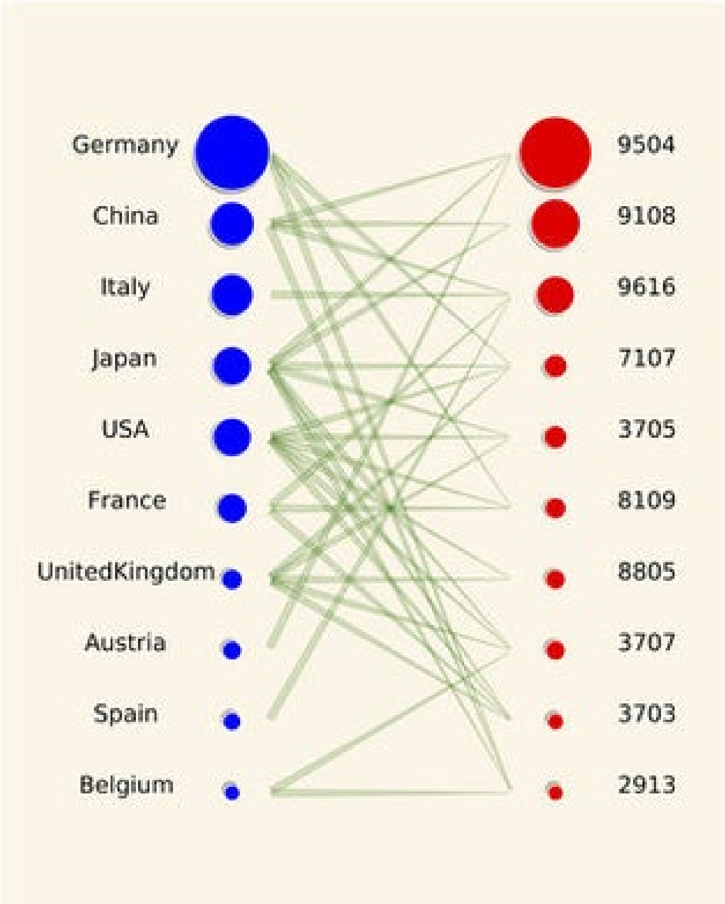
- (PRODY) The complexity of product p is the (weighted) average wealth of the countries exporting that product:

$$PRODY_p = \sum_c \frac{s_{cp}}{\sum_{c'} s_{c'p}} I_c$$

s_{cp}: share of product p in the export basket of country c;
I_c: GDP per capita (adjusted by PPP) of country c.

"Modern" measures:

- (HH - Hidalgo and Hausmann, 2009): an iterative algorithm ("method of reflections") simultaneously defining Product and Countries complexity *("the complexity of a product is the average of the complexities of the countries exporting it")*.

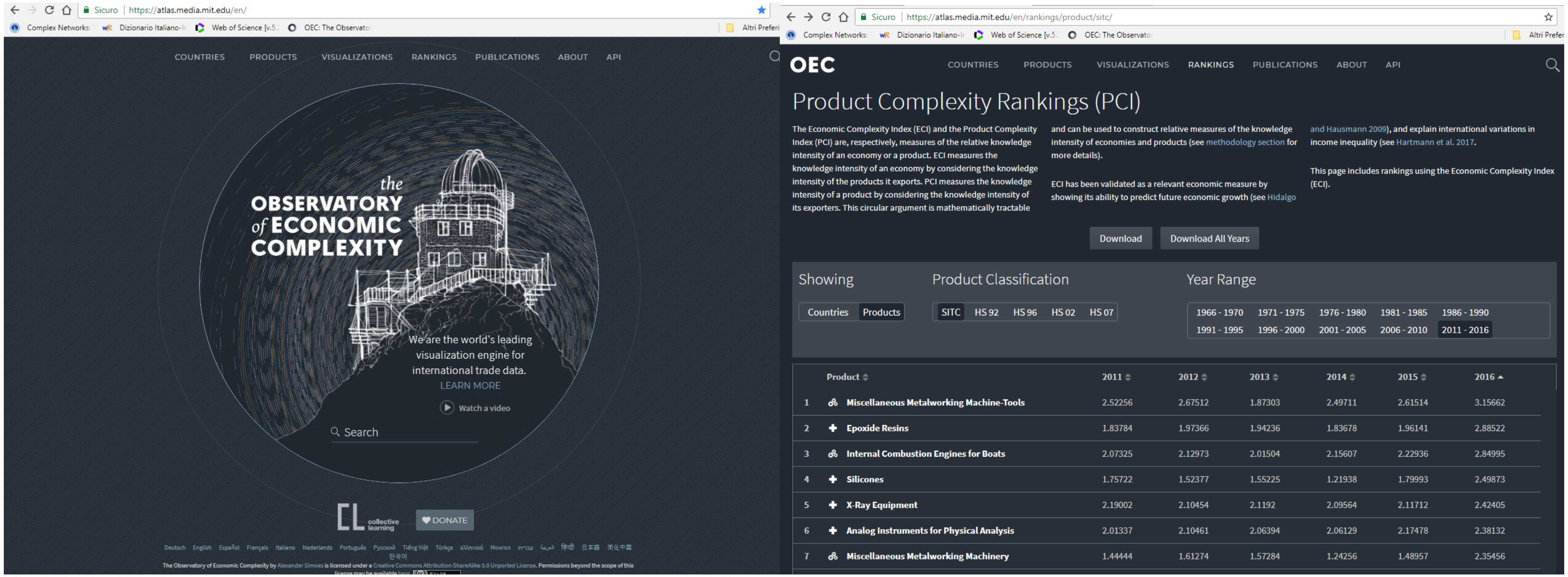


$$k_c^{(n)} = \frac{1}{k_c^{(0)}} \sum_p m_{cp} k_p^{(n-1)}$$

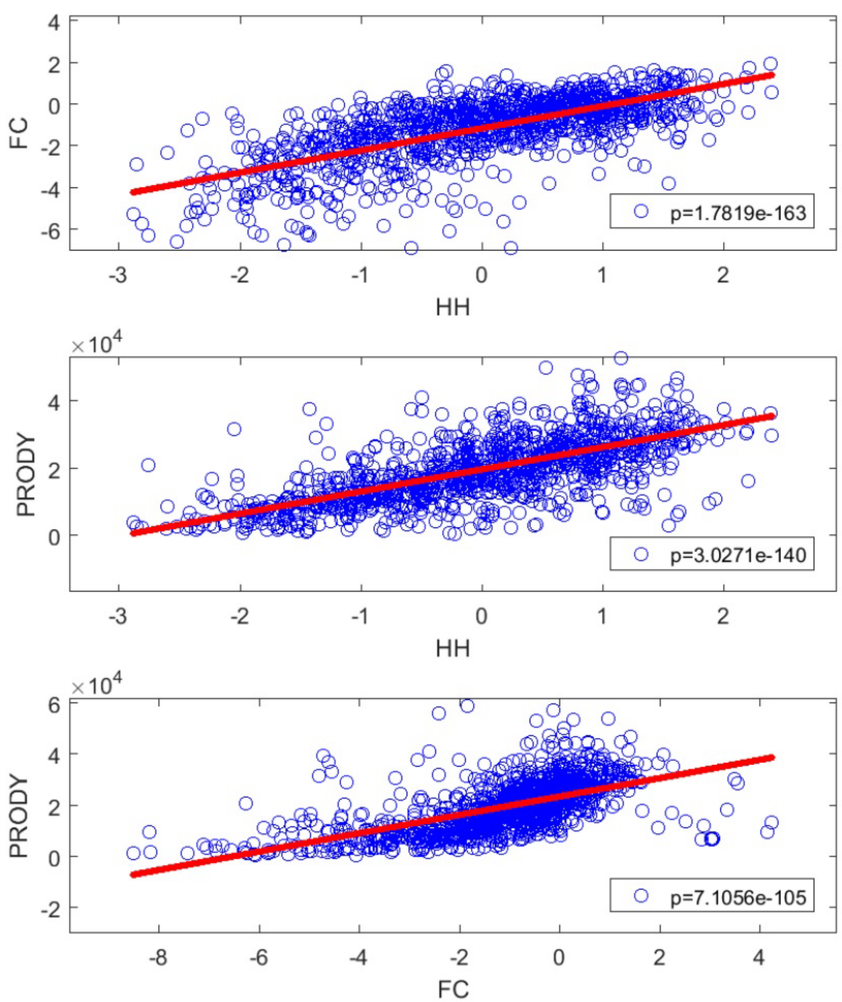
$$k_p^{(n)} = \frac{1}{k_p^{(0)}} \sum_c m_{cp} k_c^{(n-1)}$$

Rationale: Complex goods require many specific skills and inputs to be produced: their complexity can be assessed looking at the characteristics of the countries able to produce them.

Complexity values are available in a website ("the Observatory of Economic Complexity") updated yearly:



- (FC - Fitness/Complexity, Tacchella et al. 2012): a nonlinear modification of the above HH iterative algorithm, to solve some conceptual problems.

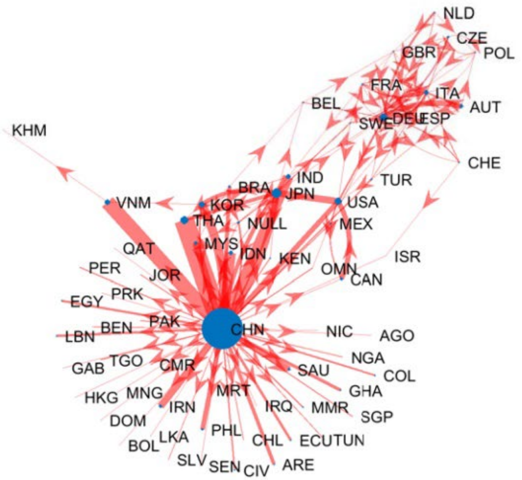


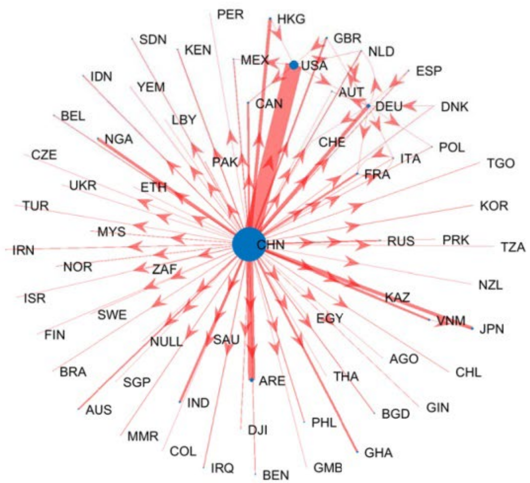
FC, HH, PRODY are highly correlated – yet remarkable differences exist for several products.

MEASURING NETWORK CENTRALIZATION

A gallery of product world trade networks:

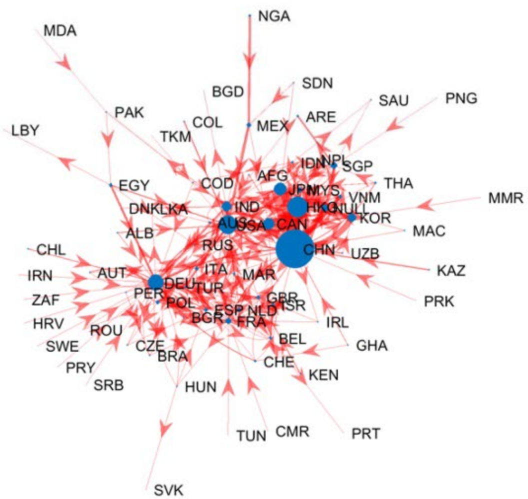
#7227
(Metal)

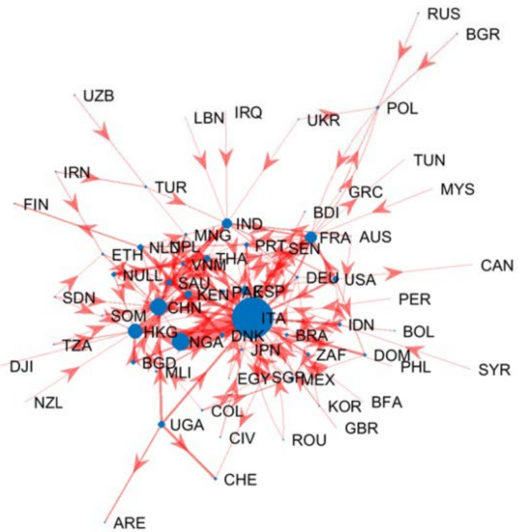




#8513
(Machin./Electr.)

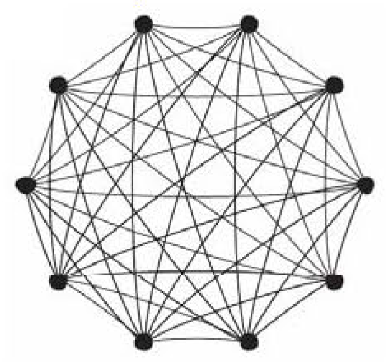
#1211
(Vegetable Prod.)





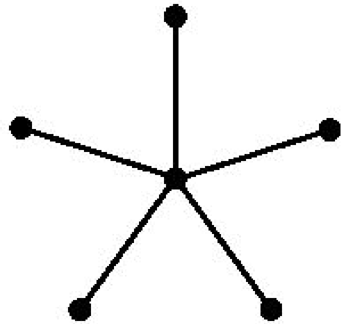
#4106
(Raw Hides,
Skins, etc)

We want to capture – product by product - the centralization of the world trade network topology:



centralization = 0

vs



centralization = 1

We use three indicators – related to topology, dynamics, and robustness.

GINI index: it directly measures the export heterogeneity:

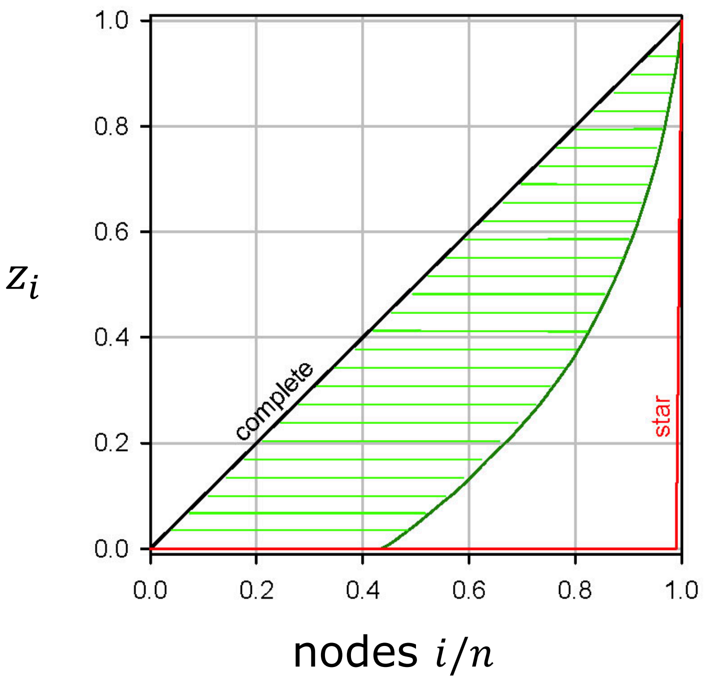
- Re-order nodes (=countries) by increasing out-strength (=export) $s_i = \sum_j w_{ij}$:

$$s_1 \leq s_2 \leq \cdots \leq s_n$$

- Define the Lorenz (cumulated) curve as

$$z_1 = s_1/S \ , \qquad z_2 = (s_1 + s_2)/S \ , \qquad z_3 = (s_1 + s_2 + s_3)/S \ , \qquad \cdots$$

where $S = \sum_i s_i = \sum_{ij} w_{ij}$ is the total world export.



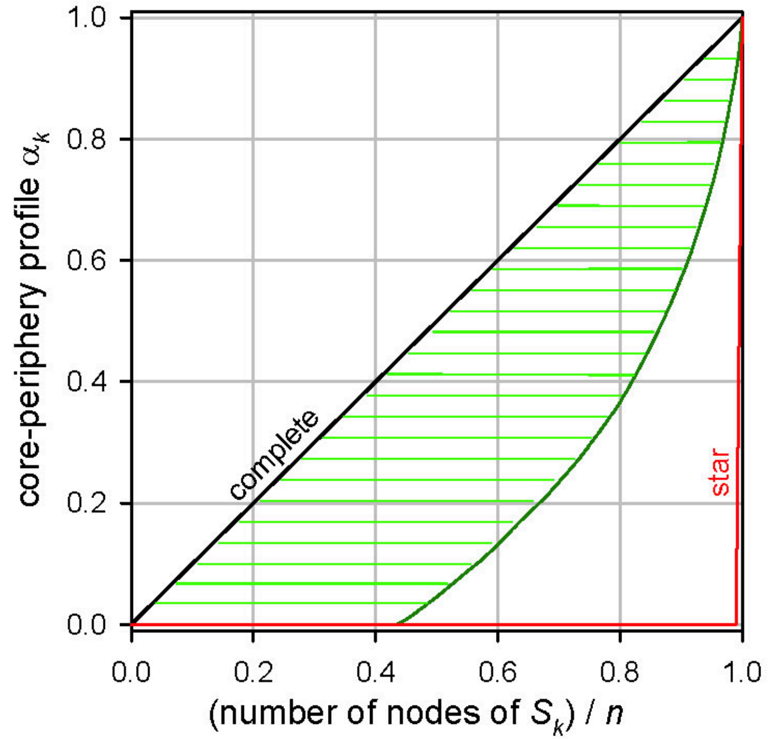
GINI is the ([0,1]-normalized) green area.

GINI index = 0: all countries have the same export.
GINI index = 1: the export is concentrated in one single country.

(CP) Core-Periphery index: it is based on the dynamics of a random-walker
[Della Rossa, Dercole, Piccardi, Sci Rep, 2013]

A heuristic procedure for ordering the nodes from to the periphery to the core:

- start by the node i with minimal strength
- generate a sequence of sets $\{i\} = S_1 \subset S_2 \subset \cdots \subset S_N = \{1,2,\ldots,N\}$ by adding, at each step, the node attaining the minimal persistence probability $\alpha_1, \alpha_2, \ldots, \alpha_N$ (=prob. that a random walker remains in S_k at the next step).



The sequence $0 = \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_N = 1$ is the Core-Periphery profile.

The CP index is the ([0,1]-normalized) green area between the Core-Periphery profile and the profile of the complete network.

(VI) Vulnerability index: how rapidly the aggregated weight is lost by node removal
[Dall'Asta, Barrat, Barthelemy, Vespignani, J Stat Mech Theory Exp, 2006]

- Re-order nodes (=countries) by decreasing out-strength (=export) $s_i = \sum_j w_{ij}$:

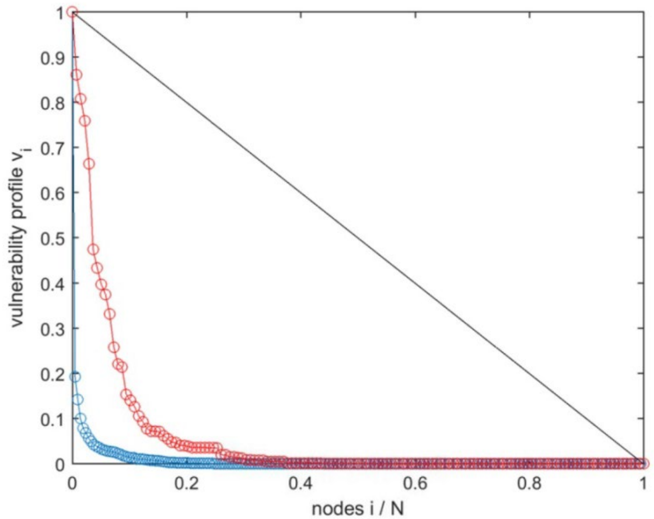
$$s_1 \geq s_2 \geq \cdots \geq s_n$$

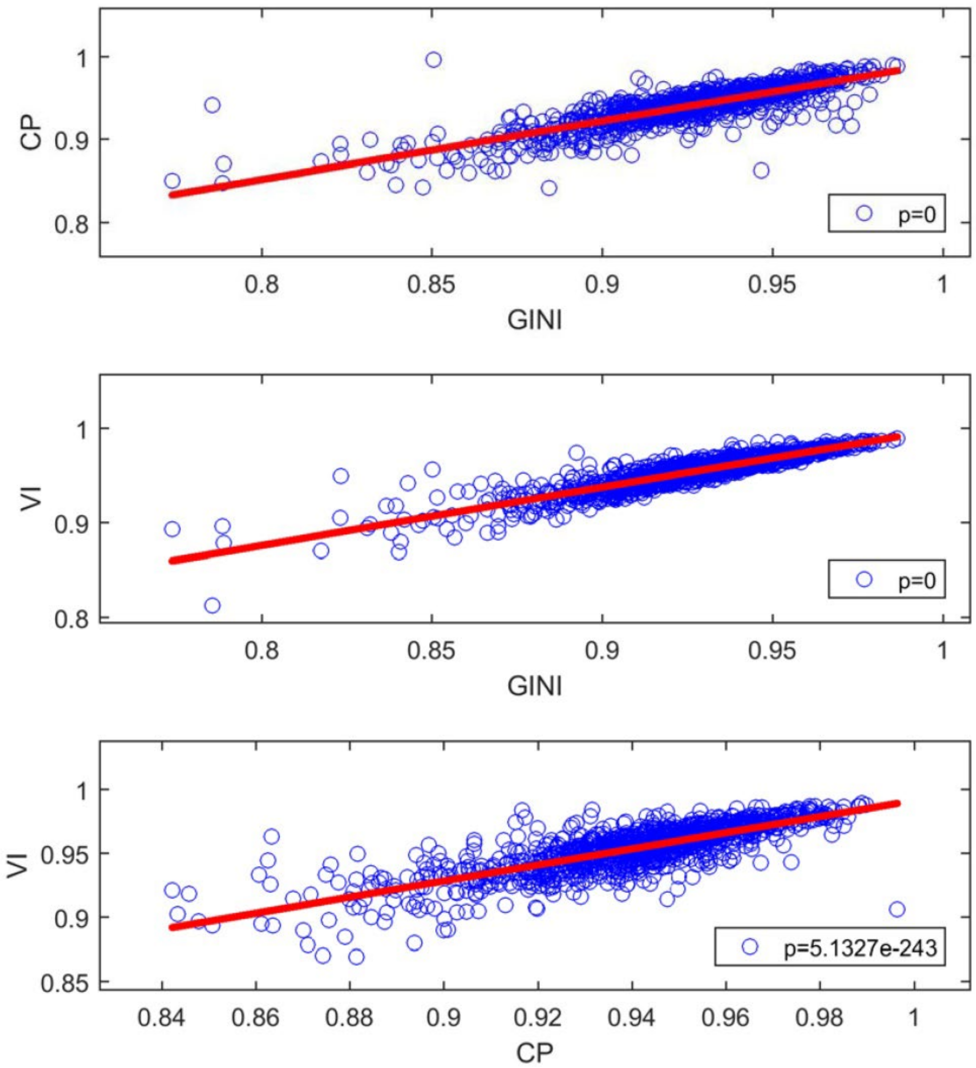
- Define the vulnerability profile as

$$1 = v_0 \geq v_1 \geq \cdots \geq v_n = 0$$

where v_k is the total network weight after (the most important) nodes $\{1,2,\cdots,k\}$ have been removed.

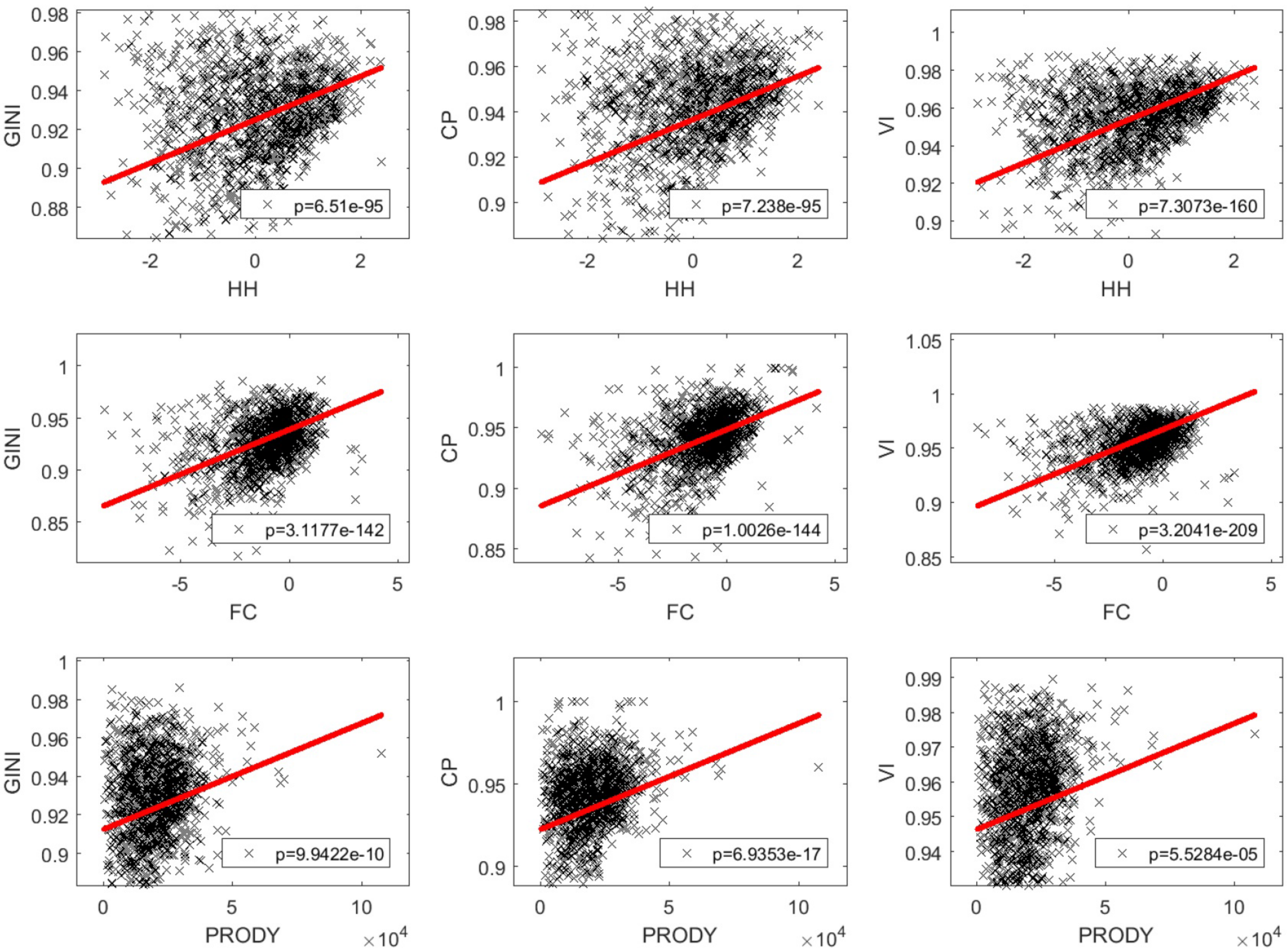
Remark: the VI index explicitly quantifies network robustness.





GINI, CP, and VI are also highly correlated.

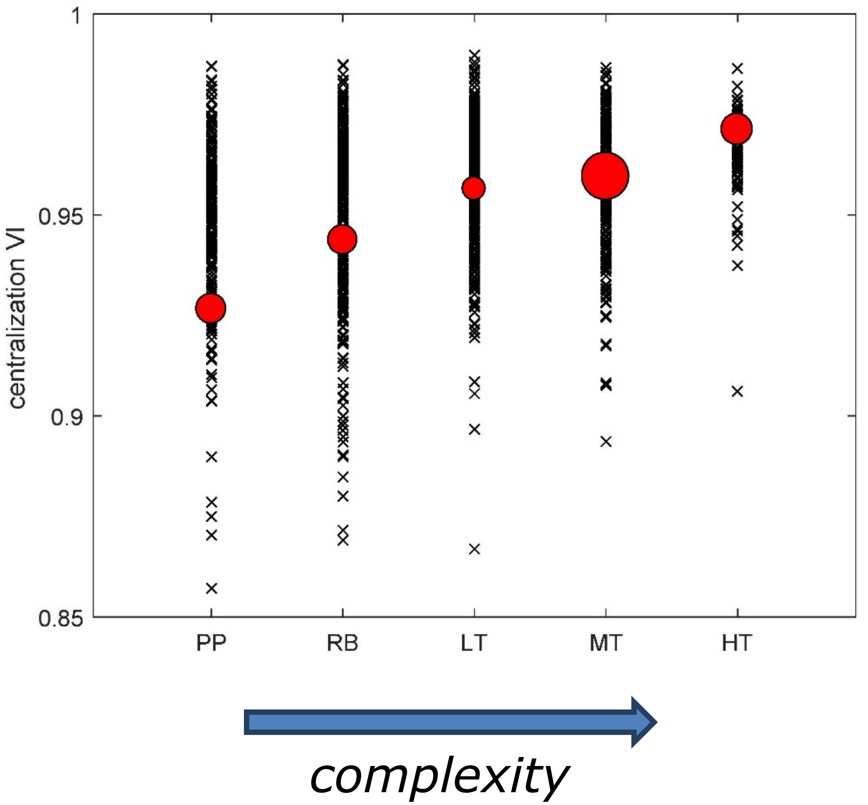
Results: Complexity vs Centralization



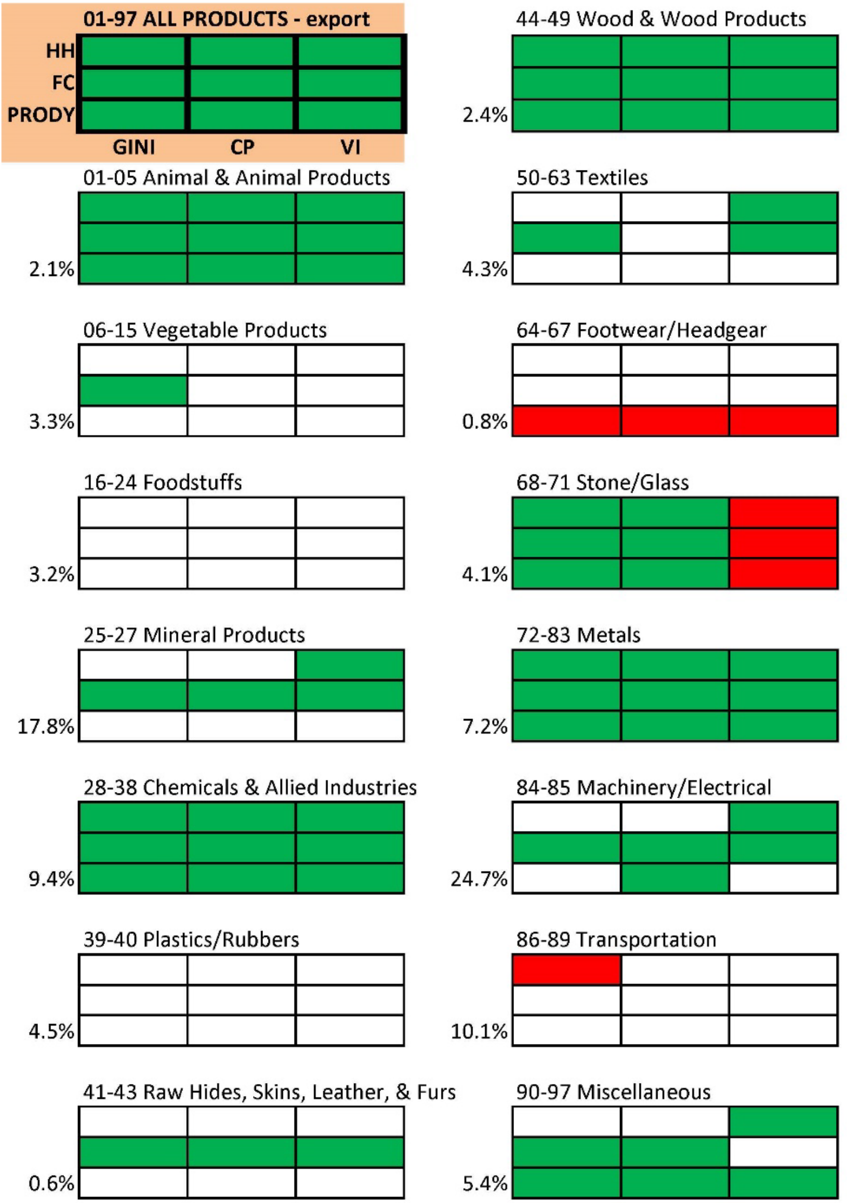
Linear regressions
(weighted by product total export)

Complexity and centralization are positively correlated
(consistently true and statistically significant for all the 3 × 3 complexity /centralization pairs.)

Cross-validation: classification based on technological class (from Primary Product – PP – to High-Technology manufacture – HT):



We have again an increasing trend of network centralization for increasing technological level.



Which categories (Sections) of products are the main drivers of the overall complexity/ centralization pattern?

We repeat the analysis on individual Sections:

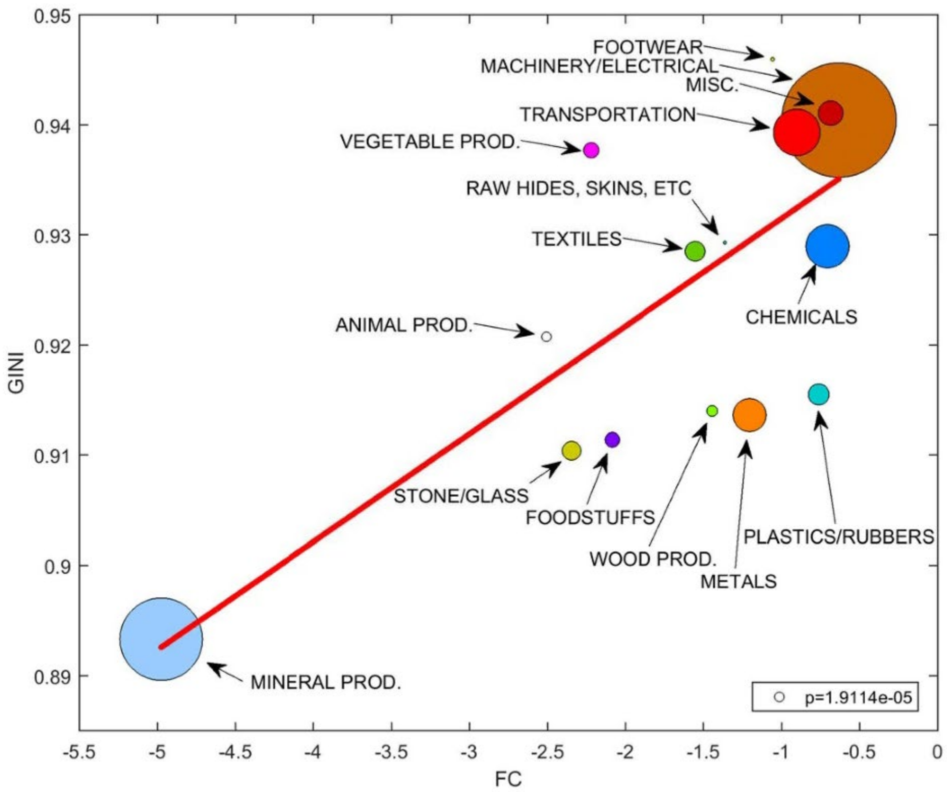
- The Sections most responsible of the overall pattern are Machinery/Electrical, Chemicals, and Metals.

- Other Sections display similar behaviour (e.g., Animal & Animal Products) but a rather small trade share.

- No Section evidences a clear opposite trend.

So far:

- Products with larger complexity are – on average – distributed through a trade network with higher centralization.
- The same holds if we separately consider some of the most important (in terms of trade volume) subsets of products.



Complementary analysis:

aggregate products by Section, and compare average complexity with average centralization.

The high-complexity=high-centralization trend is again confirmed.

- The results confirm the conjecture on the positive correlation between complexity of products and centralization of their trade networks.

- Centralization implies fragility: The more complex are the traded goods, the more fragile are their trade networks.

- Given the relevant role played by complex goods in world trade, the global trade network appears to be uncomfortably vulnerable.

PLOS | ONE

RESEARCH ARTICLE

Complexity, centralization, and fragility in economic networks

Carlo Piccardi[1]*, **Lucia Tajoli**[2]

1 Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milano, Italy,
2 Department of Management, Economics, and Industrial Engineering, Politecnico di Milano, Milano, Italy