Statistically validated networks

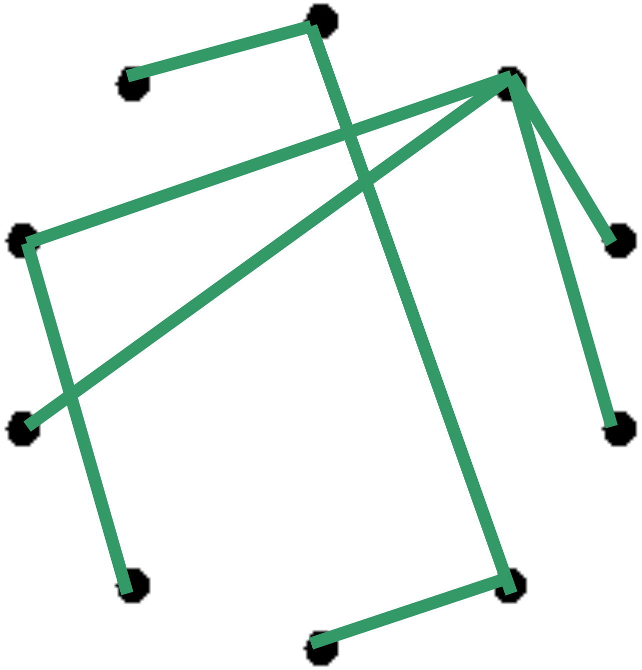
Rosario Nunzio Mantegna

Network analysis at a topology level

Let us briefly reconsider most representative network models

- Erdös-Rényi
- Scale free networks
- Small world networks
- Core periphery networks

Erdös-Rényi model (1960): Random network



Connect with probability p

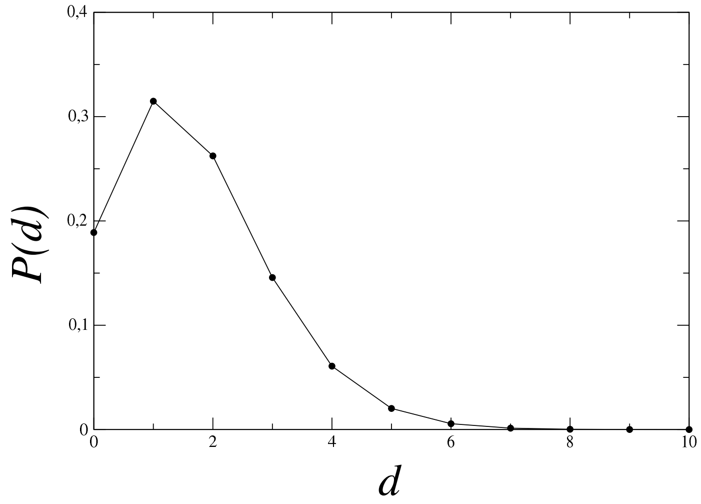
p=1/6 n=10

<*d*>=*z*=p(n-1) = 1.5

The tail of the degree distribution is decaying quickly

Poisson distribution for large n and *z*=cost

$$P(d)=\binom{n-1}{d}p^{d}(1-p)^{n-1-d}\cong\frac{(z)^{d}e^{-z}}{d!}$$



In an Erdős-Rényi network we observe an emergent phenomena: the setting of a giant component as a function of the mean degree

THE STRUCTURE AND FUNCTION OF COMPLEX NETWORKS

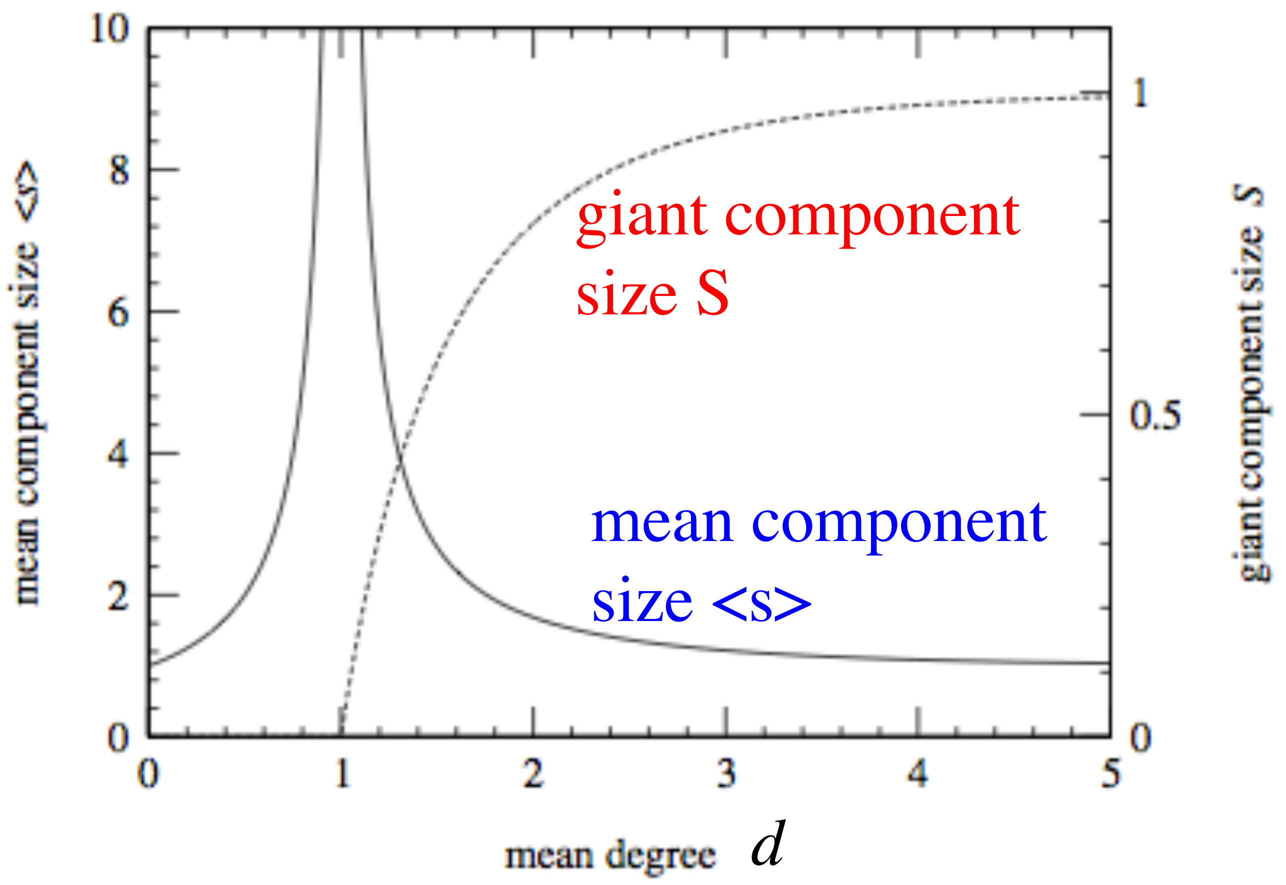
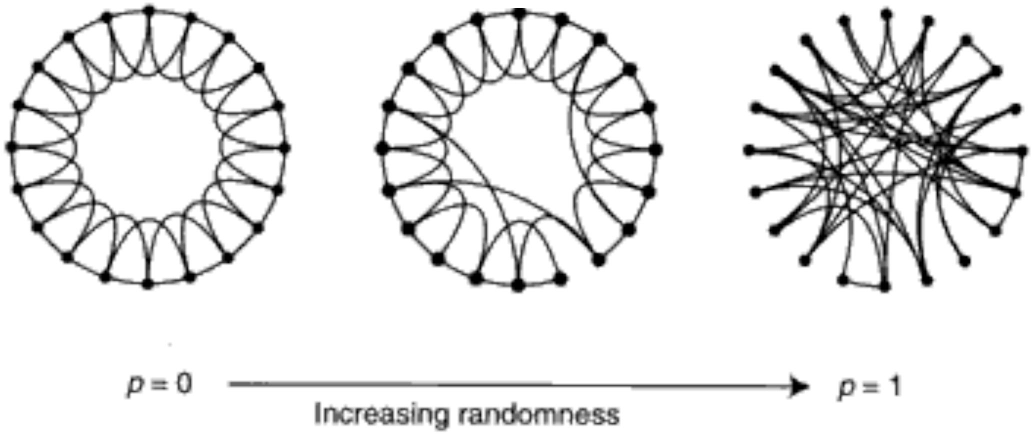


Fig. 4.1 *The mean component size (solid line), excluding the giant component if there is one, and the giant component size (dotted line), for the Poisson random graph,* (4.3) *and* (4.4).

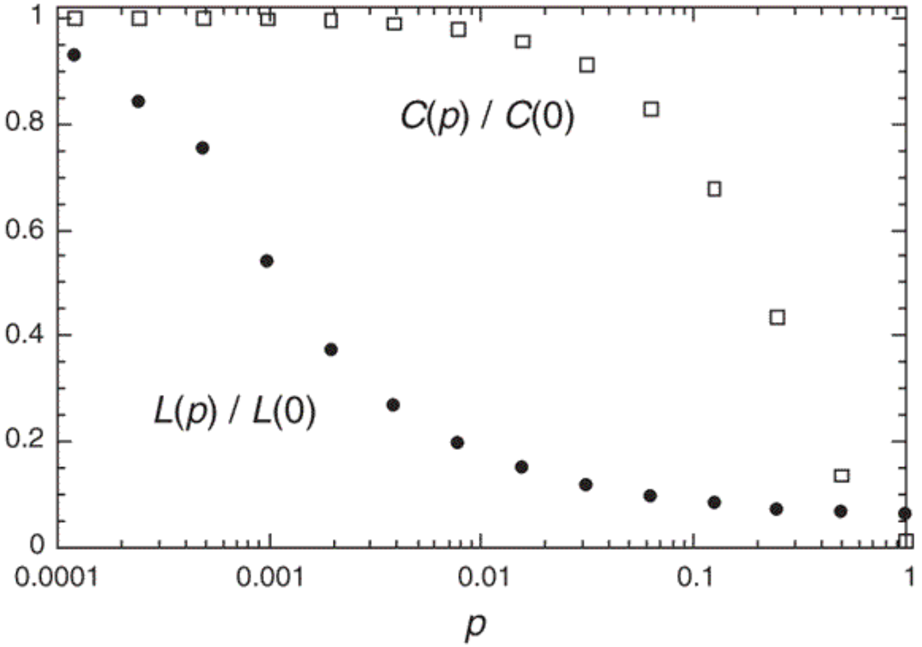
Newman, Mark EJ. "The structure and function of complex networks." SIAM review 45, no. 2 (2003): 167-256.

Watts-Strogatz model

Regular Small-world Random



Regular



p is the probability of rewiring a link

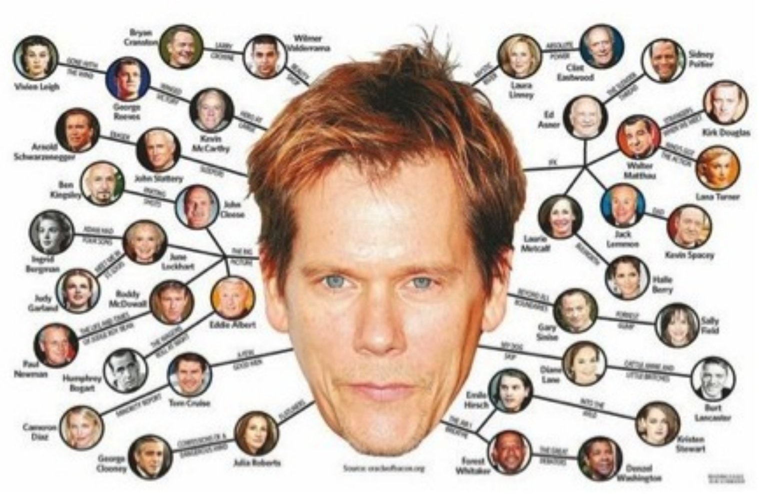
C(p) : average local clustering coeff.
L(p) : average path length

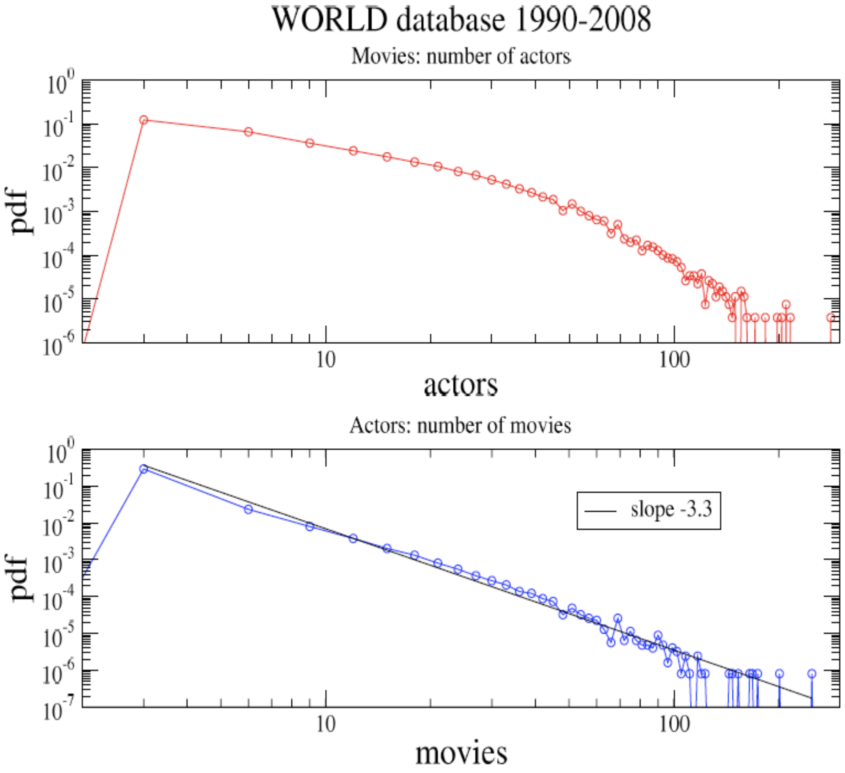
C(0) and L(0) refer to the regular lattice

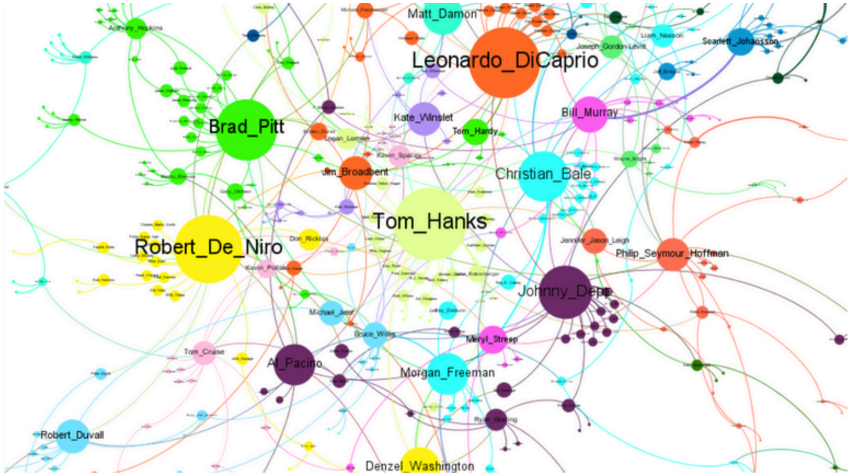
Random

Watts, Duncan J., and Steven H. Strogatz. "Collective dynamics of 'small-world'networks." Nature 393, no. 6684 (1998): 440-442.

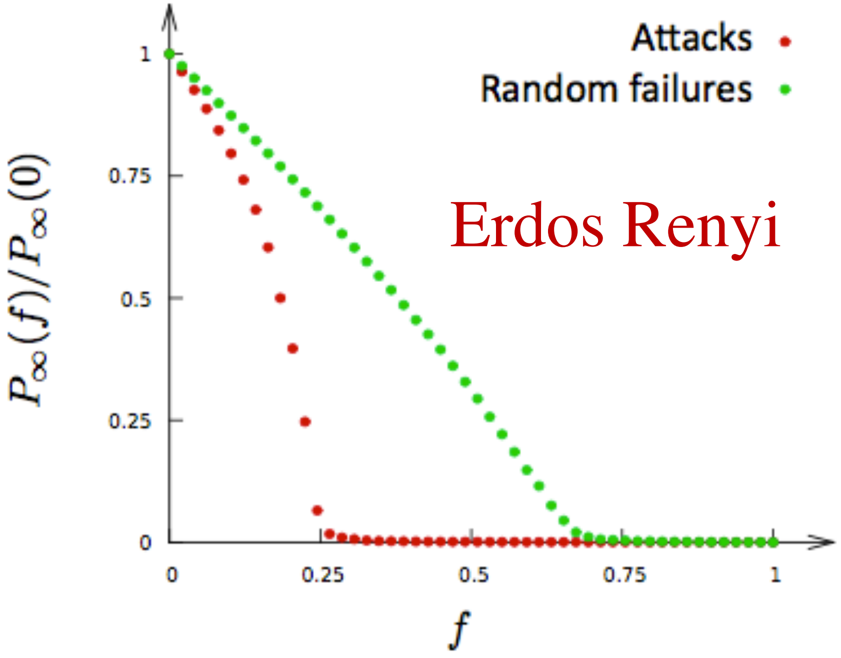
One typical characteristic of complex systems and complex network is to be heterogeneous

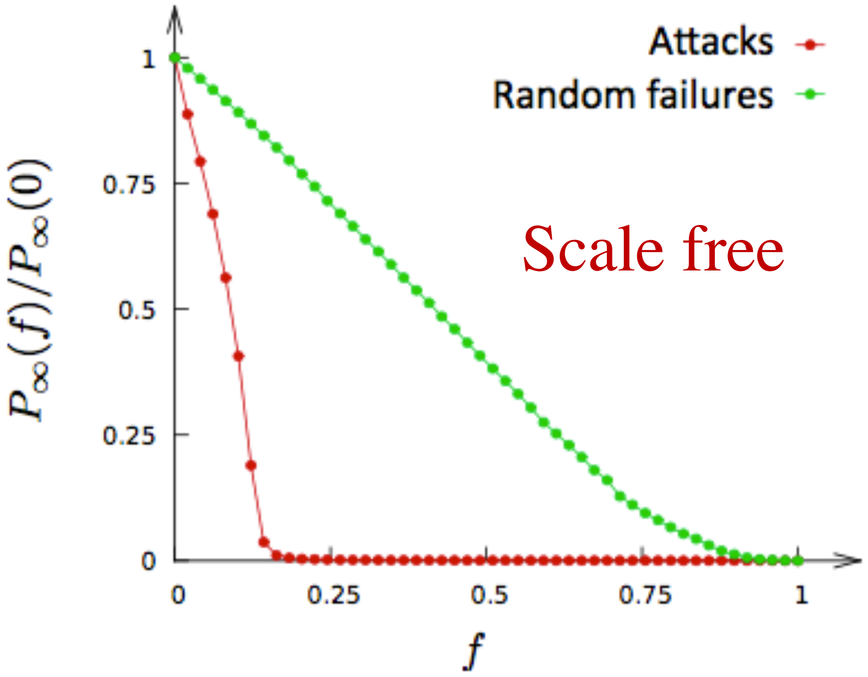






In previous lectures we saw that network topology plays an important role in characterizing network with respect to
(i) diffusion of a disease in a population,
(ii) resilience of the network to failures or attacks,

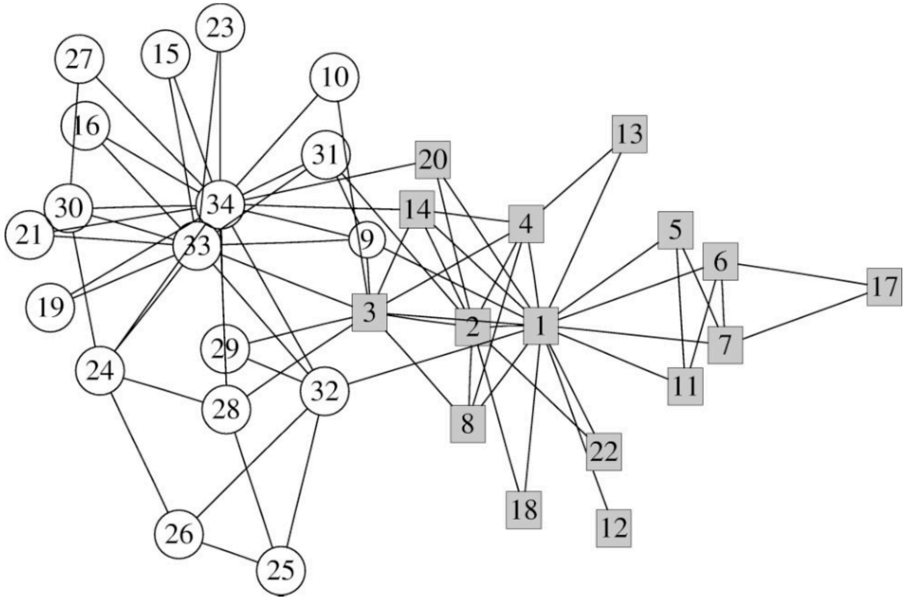




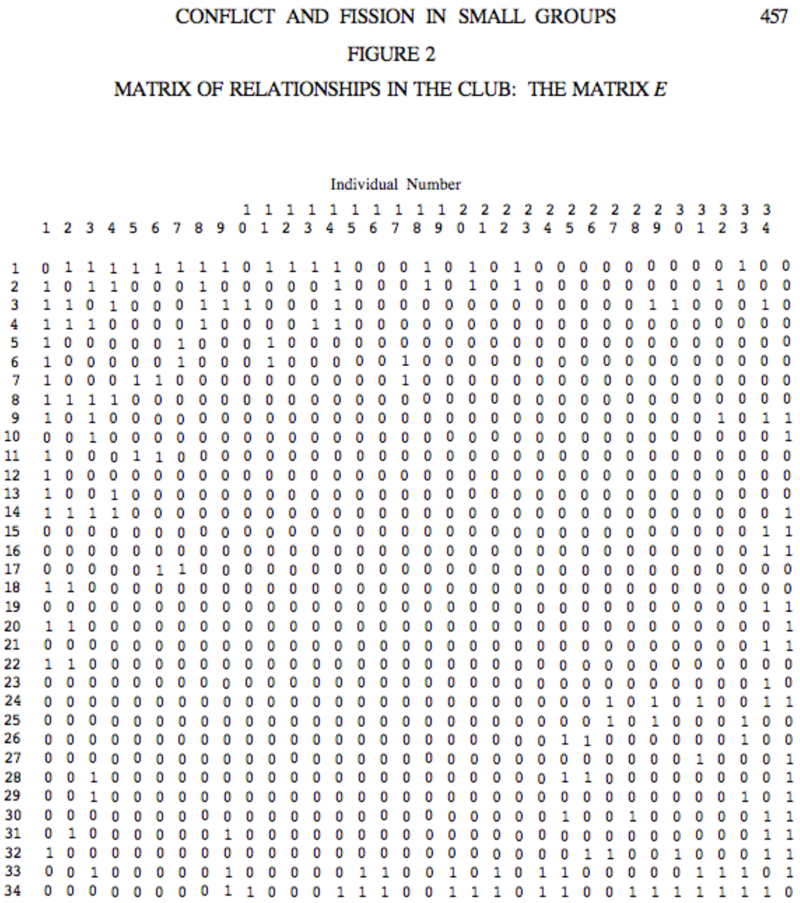
other examples are
(iii) the spread of words of mouth,
(iv) the spreading of innovation in a population, etc.

Network analysis at a community level

Another very popular type of investigation on a network is the unsupervised detection of “communities” present in it.



▶ Factions in Zachary’s karate club network. [10]



CONFLICT AND FISSION IN SMALL GROUPS 457

FIGURE 2

MATRIX OF RELATIONSHIPS IN THE CLUB: THE MATRIX *E*

Individual Number

										1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3
	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4
1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
2	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0
3	1	1	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0
4	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1
10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
11	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
17	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
20	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
22	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1	1
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
28	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1
29	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1
31	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
32	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	1	1
33	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	0	1	0	1	1	0	0	0	0	0	1	1	1	0	1
34	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1	0

Zachary, W.W., 1977. An information flow model for conflict and fission in small groups. Journal of anthropological research, 33(4), pp.452-473.

Fortunato, S. and Hric, D., 2016. Community detection in networks: A user guide. *Physics Reports*, *659*, pp.1-44.

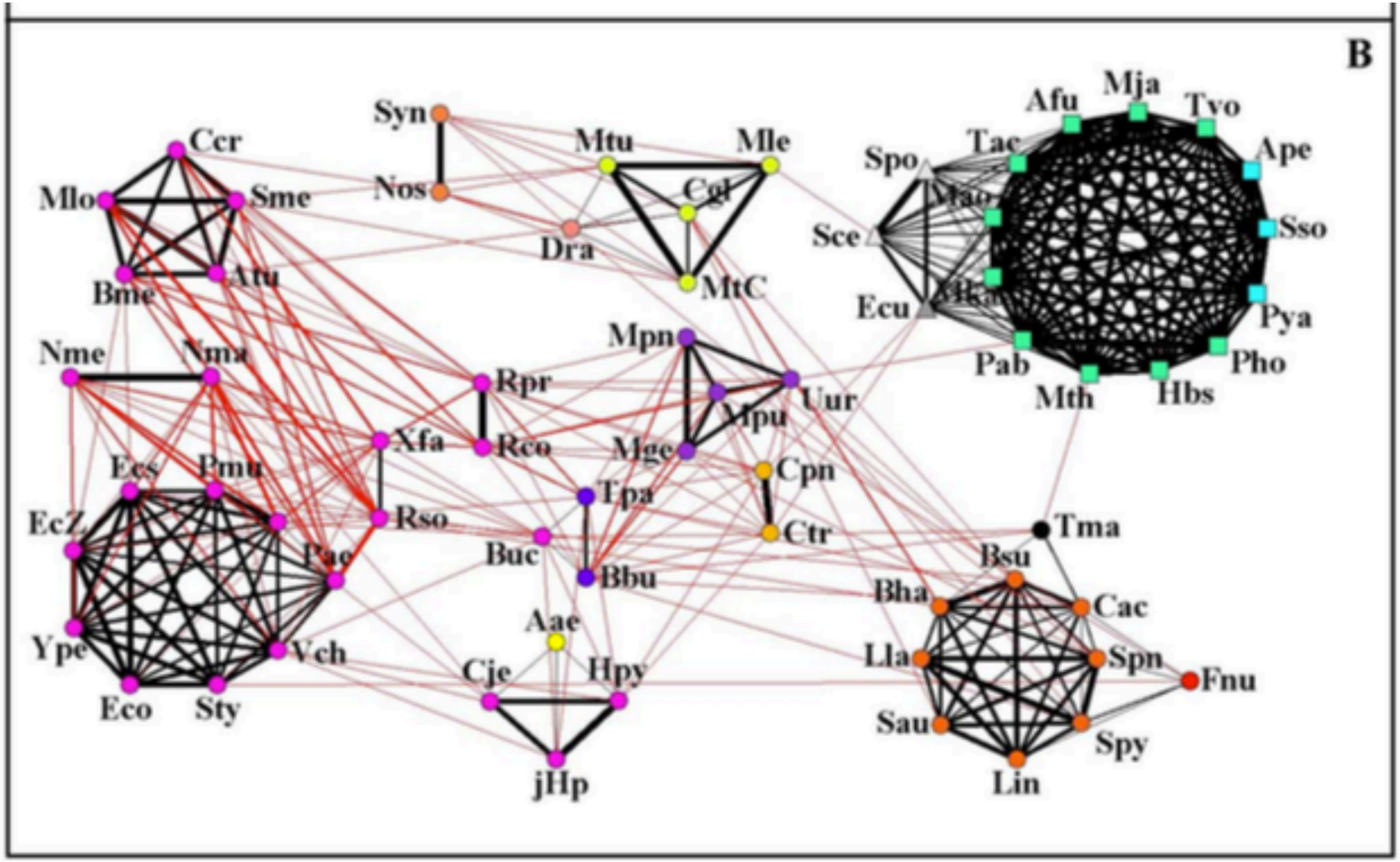


Figure 3. Statistically validated networks of organisms. Bonferroni (Panel A) and FDR (Panel B) networks of the organisms investigated in the COG database. The shape of the node indicates the super kingdom of the organism: Archaea (squares), Bacteria (circles), and Eukaryota (triangles). The color of the node indicates the phylum of the organism. The thickness of the link is related to its weight and it is proportional to the logarithm of the number of COG_k validations between the two connected nodes. Red links bridge different communities of organisms, as revealed by applying Infomap [22] to the statistically validated networks.
doi:10.1371/journal.pone.0017994.g003

Network analysis by considering network motifs

Another type of network analysis focusing on sub-networks concerns the investigation of network motifs

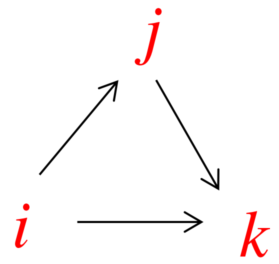
The concept of *Transitivity*

A basic concept in directed network is the concept of transitivity of a path. Most of the work on transitivity has focused on triple of actors *i*, *j*, and *k*

A list of three social actors is called a triple

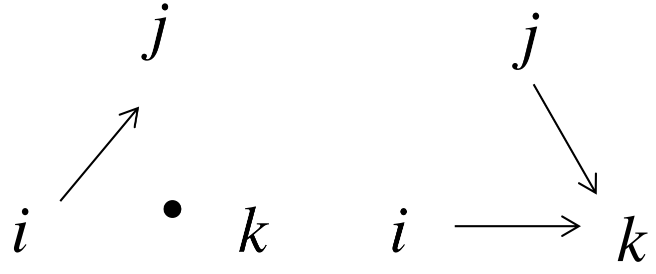
A subgraph involving the three actors and their links is called a triad

Definition: The triad involving actors i, j, and k is transitive if whenever i -> j and j -> k then i -> k

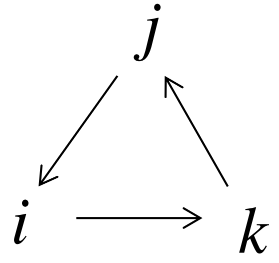


Wasserman, S. and Faust, K., 1994. Social network analysis: Methods and applications (Vol. 8). Cambridge university press.

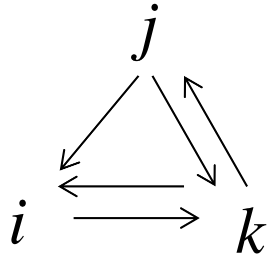
A detailed analysis of some triads with respect to the concept of transitivity shows that



Some triads are inconclusive with respect to the concept of transitivity. They are called *vacuously transitive* and are considered neither transitive nor intransitive

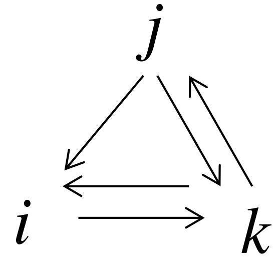


intransitive triad with one triple



intransitive triad with multiple triples

The right triad presents five links and six triples. For one triple the relation is intransitive and therefore the triad is overall considered intransitive. However, many triples of the triad are indeed transitive



Triple #1 : $n_i n_j n_k$

$n_i \circ n_j \quad n_j \to n_k \quad n_i \to n_k$

Triple #2 : $n_i n_k n_j$

$n_i \to n_k \quad n_k \to n_j \quad n_i \leftarrow n_j$

Triple #3 : $n_j n_i n_k$

$n_j \to n_i \quad n_i \to n_k \quad n_j \to n_k$

Triple #4 : $n_j n_k n_i$

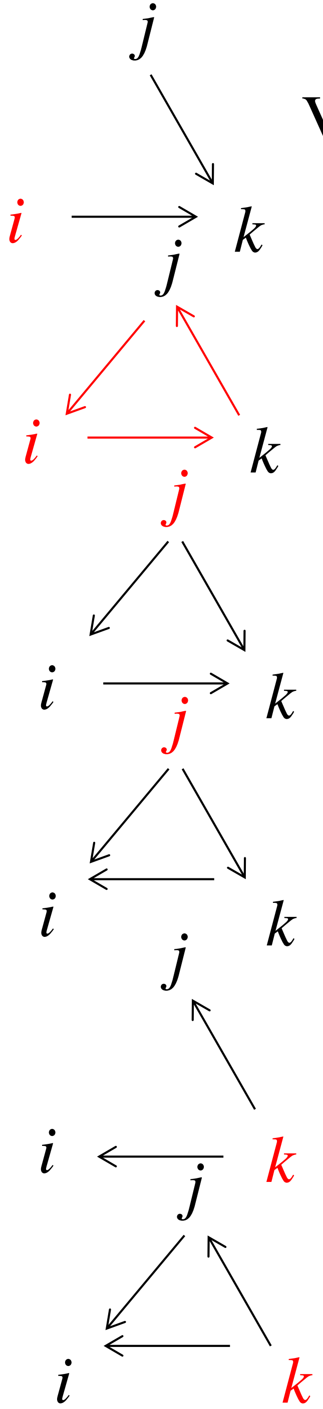
$n_j \to n_k \quad n_k \to n_i \quad n_j \to n_i$

Triple #5 : $n_k n_i n_j$

$n_k \to n_i \quad n_i \circ n_j \quad n_k \to n_j$

Triple #6 : $n_k n_j n_i$

$n_k \to n_j \quad n_j \to n_i \quad n_k \to n_i$



Vacuously transitive

Intransitive

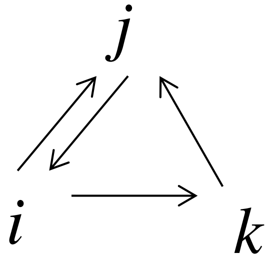
Transitive

Transitive

Vacuously transitive

Transitive

Definition of triads or 3-motifs

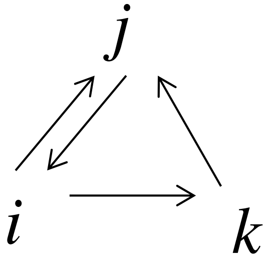


Three nodes i, j, and k can present 6 directional links among them. Each link can be present or absent. Therefore there are 2^6=64 possible triads connecting them.

However, if the specific identity of the node is not controlled, as is the case when an entire network is investigated, some of the triads are isomorphic

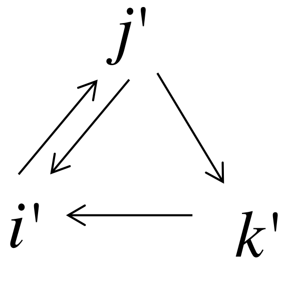
It is therefore important to detect all classes of isomorphic triads.

For example the triad



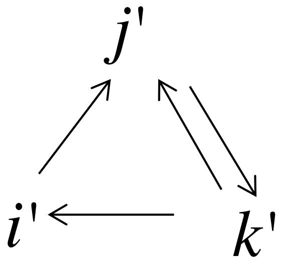
	i	j	k
i	0	1	1
j	1	0	0
k	0	1	0

is isomorphic to



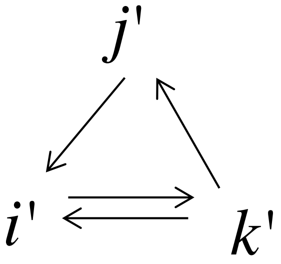
	i'	j'	k'
i'	0	1	0
j'	1	0	1
k'	1	0	0

k -> k' ; i -> j' ; j -> i'

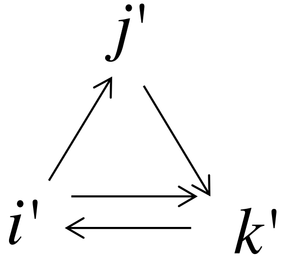


	i'	j'	k'
i'	0	1	0
j'	0	0	1
k'	1	1	0

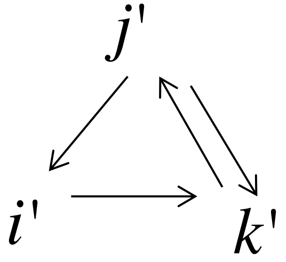
k -> i' ; i -> k' ; j -> j'



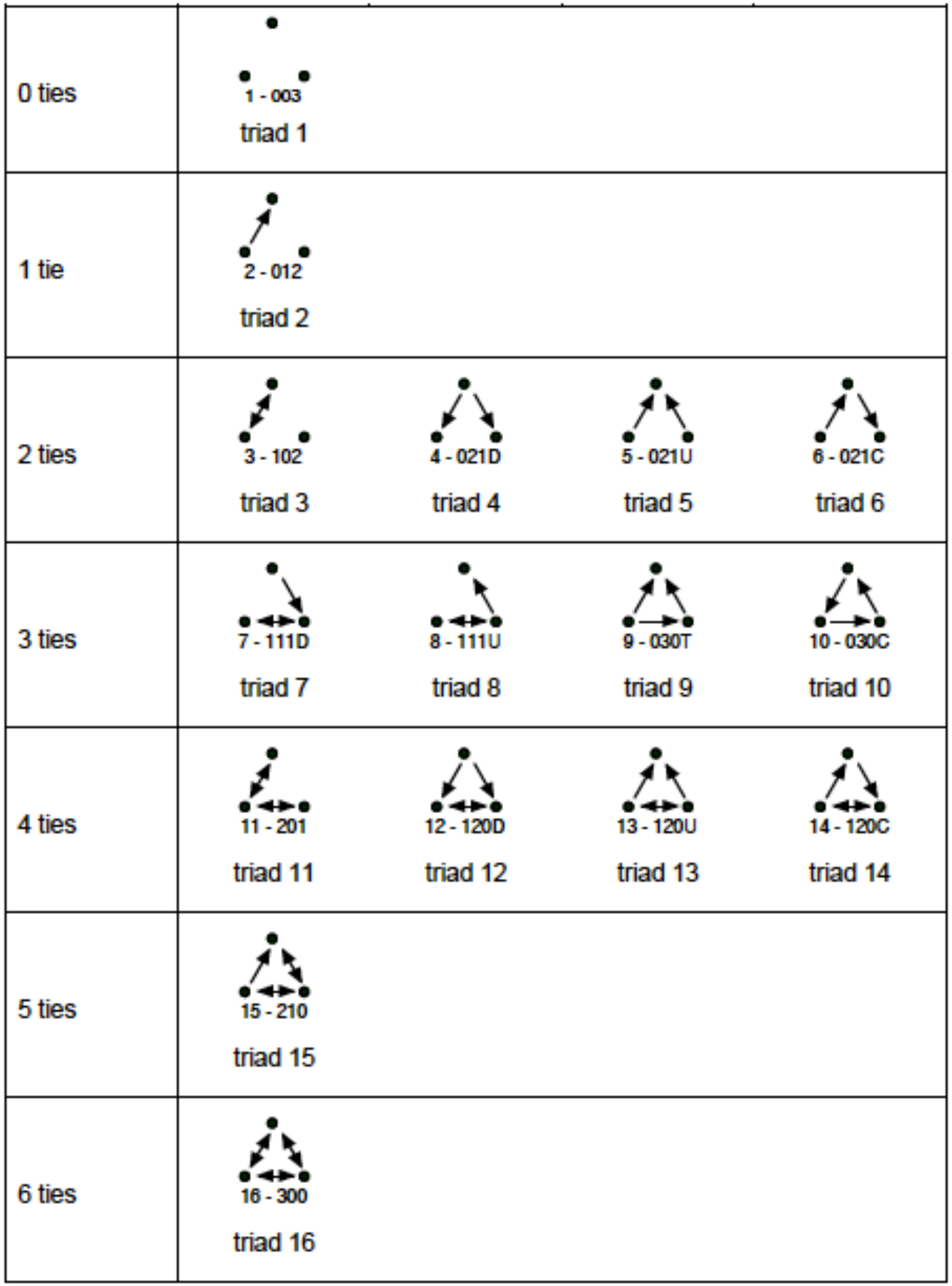
	i'	j'	k'
i'	0	0	1
j'	1	0	0
k'	1	1	0



	i'	j'	k'
i'	0	1	1
j'	0	0	1
k'	1	0	0



	i'	j'	k'
i'	0	0	1
j'	1	0	1
k'	0	1	0



0 ties	1 - 003 triad 1			
1 tie	2 - 012 triad 2			
2 ties	3 - 102 triad 3	4 - 021D triad 4	5 - 021U triad 5	6 - 021C triad 6
3 ties	7 - 111D triad 7	8 - 111U triad 8	9 - 030T triad 9	10 - 030C triad 10
4 ties	11 - 201 triad 11	12 - 120D triad 12	13 - 120U triad 13	14 - 120C triad 14
5 ties	15 - 210 triad 15			
6 ties	16 - 300 triad 16			

The sixteen classes of isomorphic triads ordered. Davis and Leinhardt (1968,1972)

A classic example: The study of Krackhardt's friendship relation among high-tech managers.

N_V=21 There are $N_V(N_V-1)(N_V-2)/6=1330$ triads

The triad census is

$T_{friends}=(376,366,143,114,34,35,39,101,23,0,20,16,25,9,23,6)$

Krackhardt, D., 1987. Cognitive social structures. Social networks, 9(2), pp.109-134.

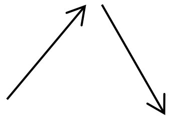
The number of triads observed has to be compared with the ones expected for a null hypothesis.

The ingredients typically maintained in the random null hypothesis are:
- in degree distribution;
- out degree distribution;
- number of mutual links present in the system.

Triad type	Triad census	Expected value	Standard deviation	z-score
003	376	320.06	9.39	5.96
012	366	416.82	14.56	-3.49
102	143	171.19	9.43	-2.99
021D	114	44.09	6.22	11.24
021U	34	44.09	6.22	-1.62
021C	35	88.17	8.17	-6.51
111D	39	73.74	7.78	-4.47
111U	101	73.74	7.78	3.50
030T	23	18.17	3.86	1.25
030C	0	6.06	2.39	-2.54
201	20	28.97	4.52	-1.98
120D	16	7.74	2.71	3.05
120U	25	7.74	2.71	6.37
120C	9	15.48	3.74	-1.73
210	23	12.38	3.25	3.27
300	6	1.55	1.2	3.71

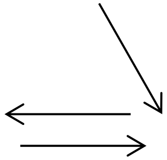
$$z = \frac{x - \mathrm{E}(x)}{\sigma_x}$$

Top 4 under expressed triads



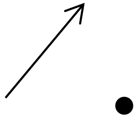
021C

z=-6.51



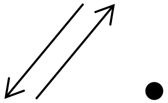
111D

z=-4.47



012

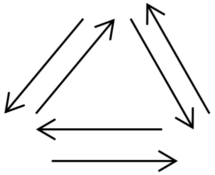
z=-3.49



102

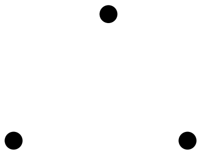
z=-3.49

Top 4 over expressed triads



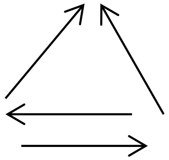
300

z=+3.71



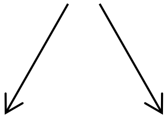
003

z=+5.96



120U

z=+6.37



021D

z=+11.2

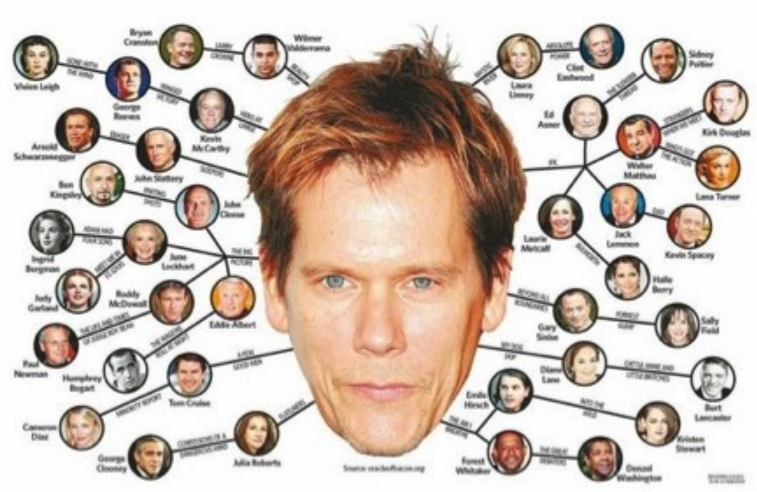
Network analysis by considering each single link

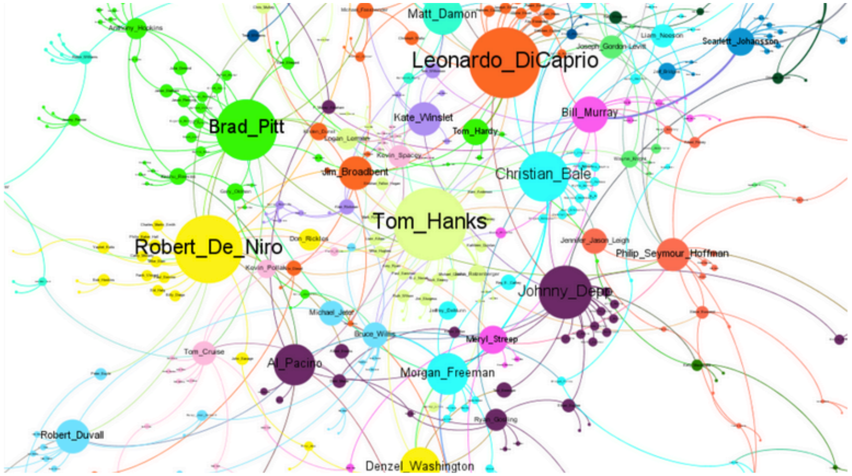
One can also go down to the level of statistically comparing a single link with a given null hypothesis.

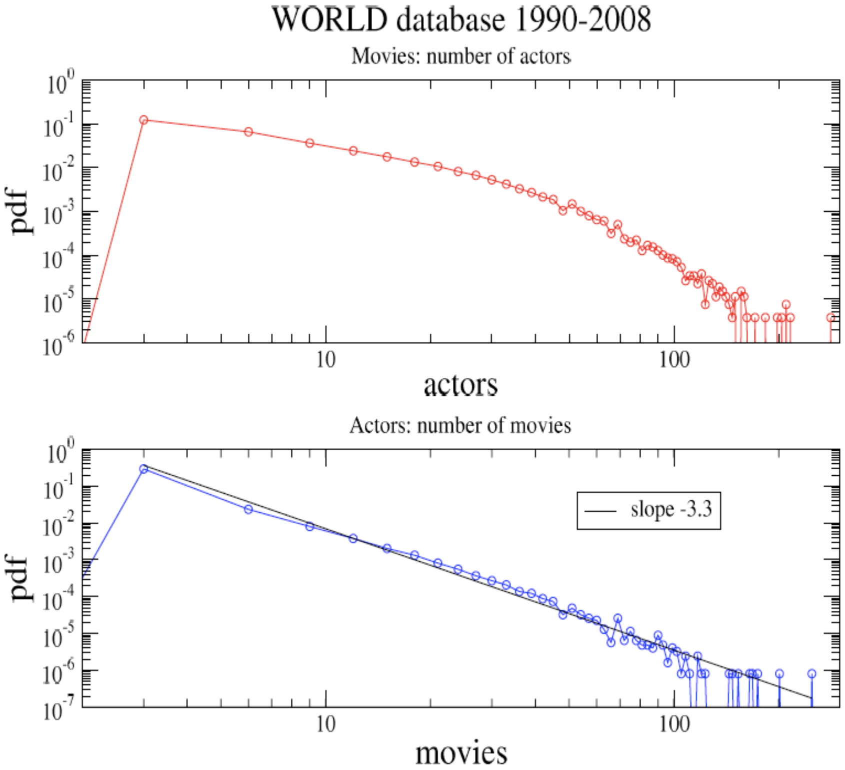
This can be done to highlight “significant” links.

- detect the "backbone" of a network;
- filter a highly dense (sometime fully connected) network.
- perform pre-processing of datasets in complex databases.
- highlight cores in community detection

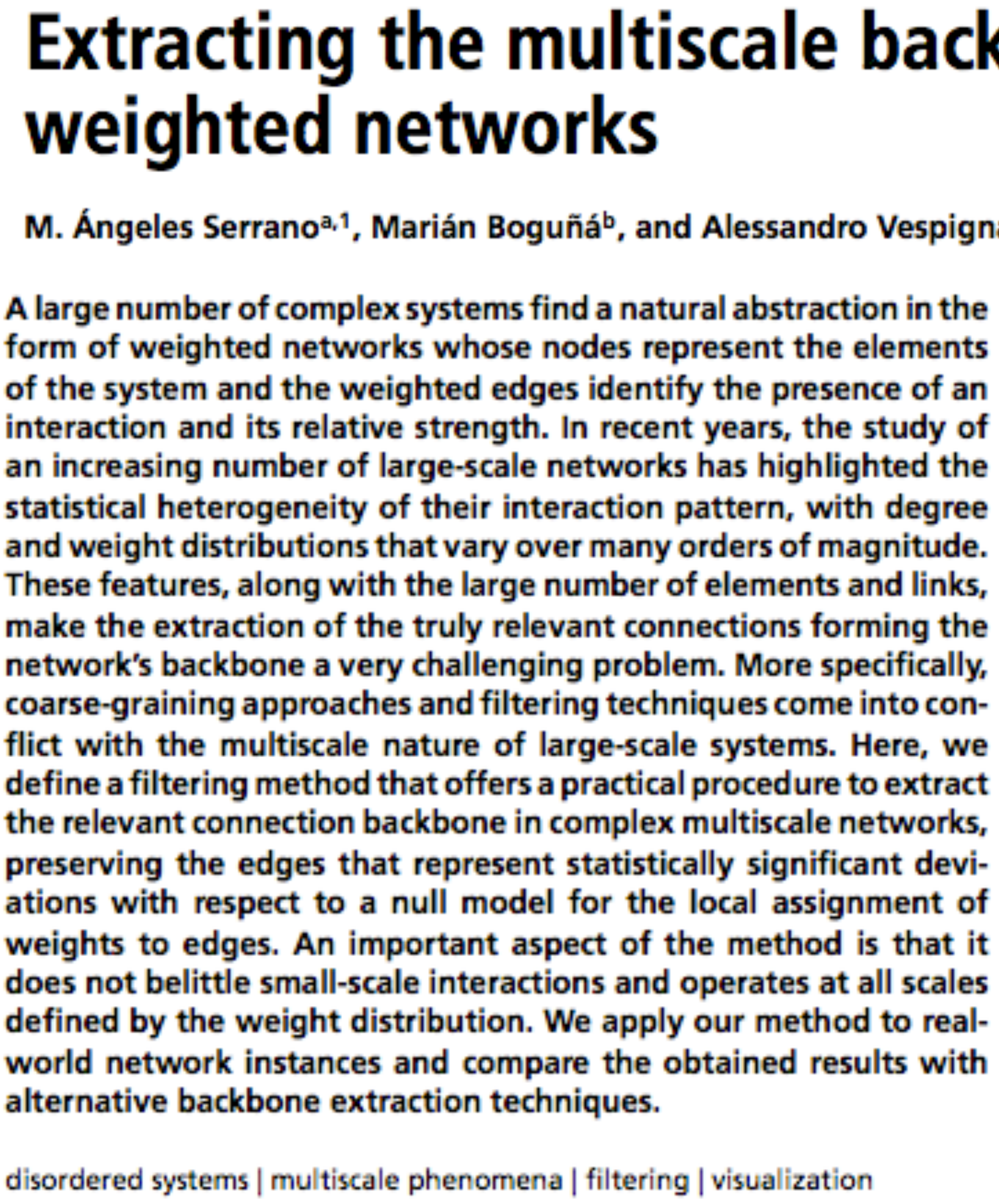
One typical characteristic of complex systems and complex network is to be heterogeneous







A filtering method for weighted networks



Extracting the multiscale backbone of complex weighted networks

M. Ángeles Serrano[a,1], Marián Boguñá[b], and Alessandro Vespignani[c,d]

A large number of complex systems find a natural abstraction in the form of weighted networks whose nodes represent the elements of the system and the weighted edges identify the presence of an interaction and its relative strength. In recent years, the study of an increasing number of large-scale networks has highlighted the statistical heterogeneity of their interaction pattern, with degree and weight distributions that vary over many orders of magnitude. These features, along with the large number of elements and links, make the extraction of the truly relevant connections forming the network's backbone a very challenging problem. More specifically, coarse-graining approaches and filtering techniques come into conflict with the multiscale nature of large-scale systems. Here, we define a filtering method that offers a practical procedure to extract the relevant connection backbone in complex multiscale networks, preserving the edges that represent statistically significant deviations with respect to a null model for the local assignment of weights to edges. An important aspect of the method is that it does not belittle small-scale interactions and operates at all scales defined by the weight distribution. We apply our method to real-world network instances and compare the obtained results with alternative backbone extraction techniques.

disordered systems | multiscale phenomena | filtering | visualization

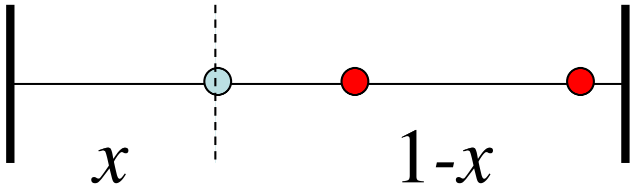
Serrano, M. Ángeles, Marián Boguná, and Alessandro Vespignani. "Extracting the multiscale backbone of complex weighted networks." Proceedings of the national academy of sciences 106, no. 16 (2009): 6483-6488.

The disparity filter

For each node i a null hypothesis is stated. The null hypothesis assumes that the strength of node i is distributed uniformly across the k_i nodes of its k_i first neighbors.

Under this null hypothesis the probability that a generic link of node i will have associated a fraction of weight x is

$$\rho(x)dx = (k-1)(1-x)^{k-2}dx$$



example with k=4

The probability of not observing k-2 breaks in an interval covering x fraction of the interval is given by the binomial probability

$$x^0(1-x)^{k-2}$$

By using the probability density associated with this null hypothesis it is possible to obtain a p-value p(x_{ij}) for each fraction of the strength x_{ij} of an edge $\{i,j\}$

$$p\left(x_{ij}\right) = 1 - \left(k-1\right)\int_{0}^{x_{ij}}\left(1-x\right)^{k-2}dx$$

The filtering of the node is then performed when the p-value $p(x_{ij})$ is less than a pre-defined statistical threshold α

The filter is providing directed links because the test is done on the fraction of strength of node i and this implies that, in general $p(x_{ij}) \neq p(x_{ji})$

In their work authors do not perform multiple hypothesis test correction.

It should be noted that this procedure is different from just performing a filtering of the links with a small fraction of strength by using a general global threshold.

The reason why the use of a general global threshold is discouraged is the fact that complex systems are highly heterogeneous and they are rarely described by a finite set of scales.

In fact simple by using standard indicators of concentration, such as the Herfindahl-Hirschman index

$$\gamma_i(k) = k_i \sum_j x_{ij}^2$$

one is able to assess the effect of inhomogeneity in the weights of links of node *i*

Authors applied their filtering method to two networks widely investigated in network science

US airports network
Florida Bay Food Web

Basic properties of the backbone at different values of the statistical threshold α

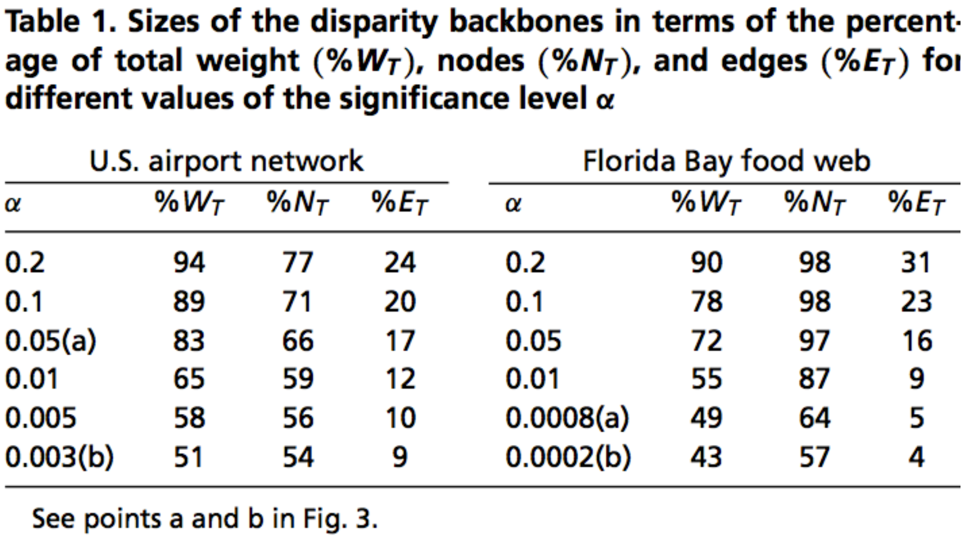


Table 1. Sizes of the disparity backbones in terms of the percentage of total weight (%W_T), nodes (%N_T), and edges (%E_T) for different values of the significance level α

U.S. airport network				Florida Bay food web			
α	%W_T	%N_T	%E_T	α	%W_T	%N_T	%E_T
0.2	94	77	24	0.2	90	98	31
0.1	89	71	20	0.1	78	98	23
0.05(a)	83	66	17	0.05	72	97	16
0.01	65	59	12	0.01	55	87	9
0.005	58	56	10	0.0008(a)	49	64	5
0.003(b)	51	54	9	0.0002(b)	43	57	4

See points a and b in Fig. 3.

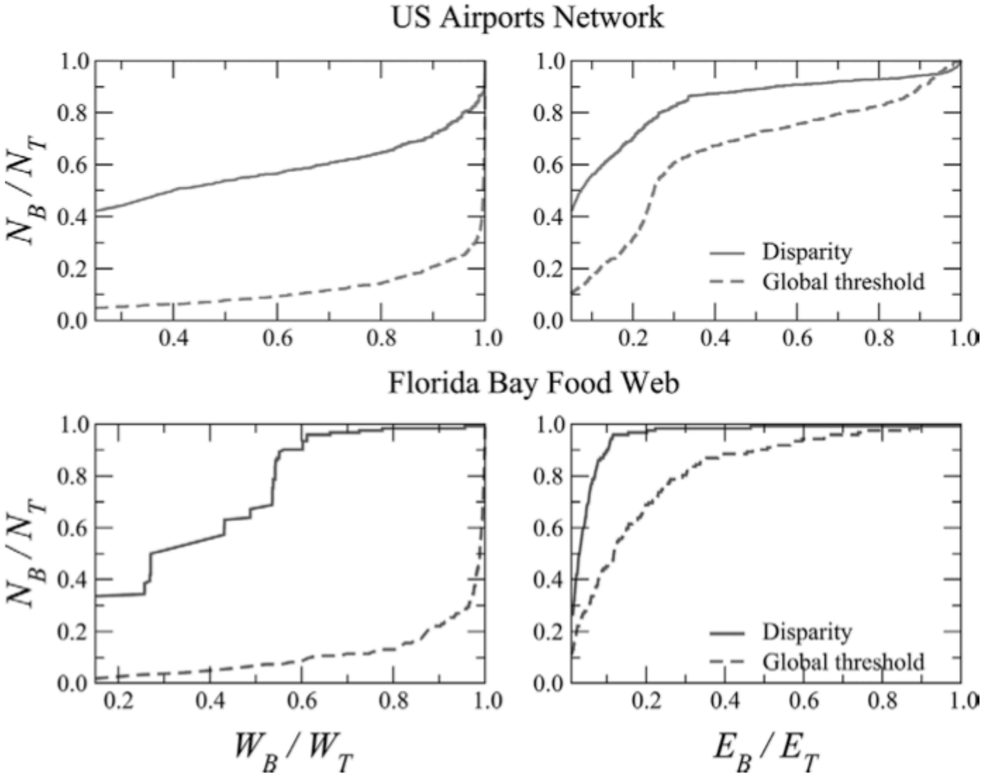


Fig. 1. Fraction of nodes kept in the backbones as a function of the fraction of weight (*Left*) and edges (*Right*) retained by the filters.

Cumulative degree distribution

Comparison of the disparity filter and a global threshold filter (removing all links with a fraction of strength below W_B)

Density function of links' weights

Clustering coefficient averaged over nodes of degree >1

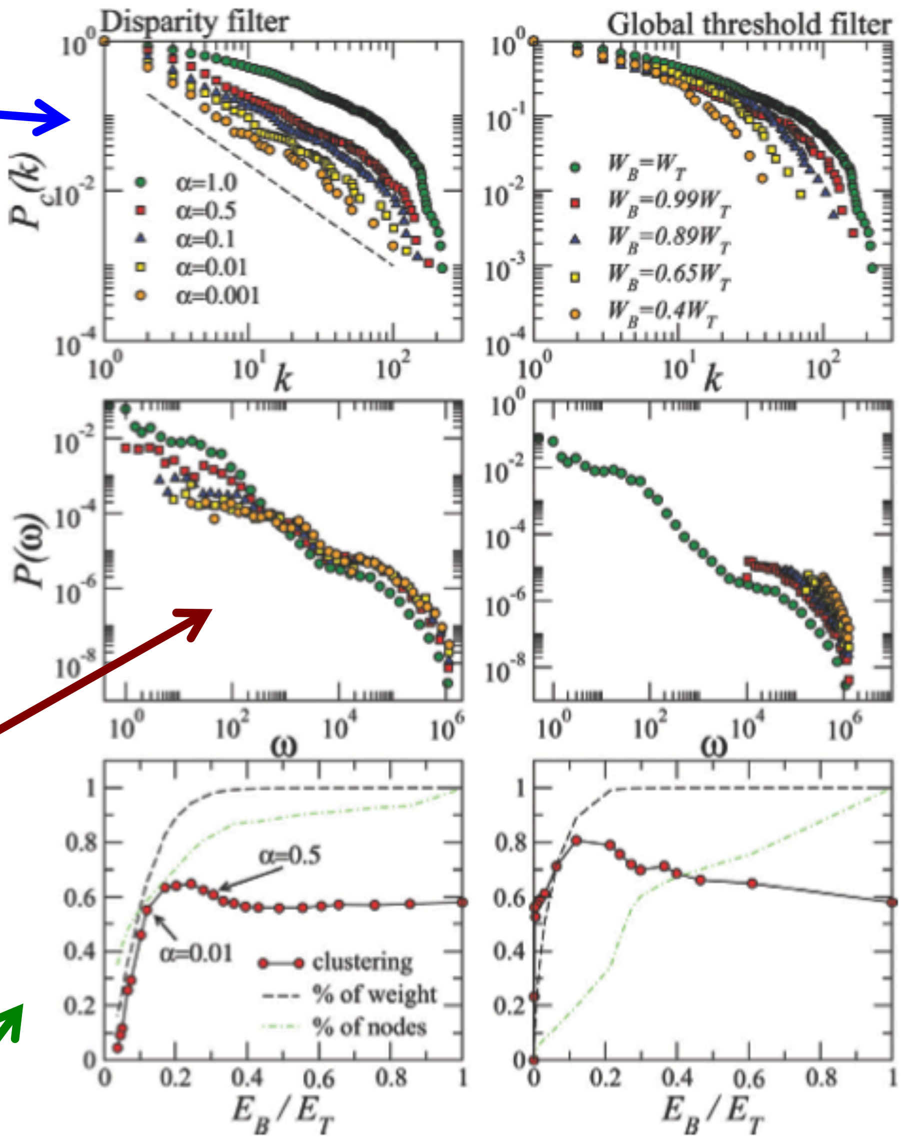


Fig. 2. Topology of the filtered subgraphs for the U.S. airports network. (*Top*) Cumulative degree distribution, $P_c(k)$, for the disparity (*Left*) and global threshold (*Right*) backbones. The values of ω_c on the right plot are chosen to generate subgraphs with the same weight as the ones shown on the left plot. (*Middle*) Distribution of links' weights of the different subgraphs generated by the two filters. Symbols are the same as in the top plots. (*Bottom*) Clustering coefficient averaged over nodes of degrees >1 for the two methods as a function of the fraction of edges in the backbones. Dashed lines show the fraction of nodes and weight for a given fraction of edges.

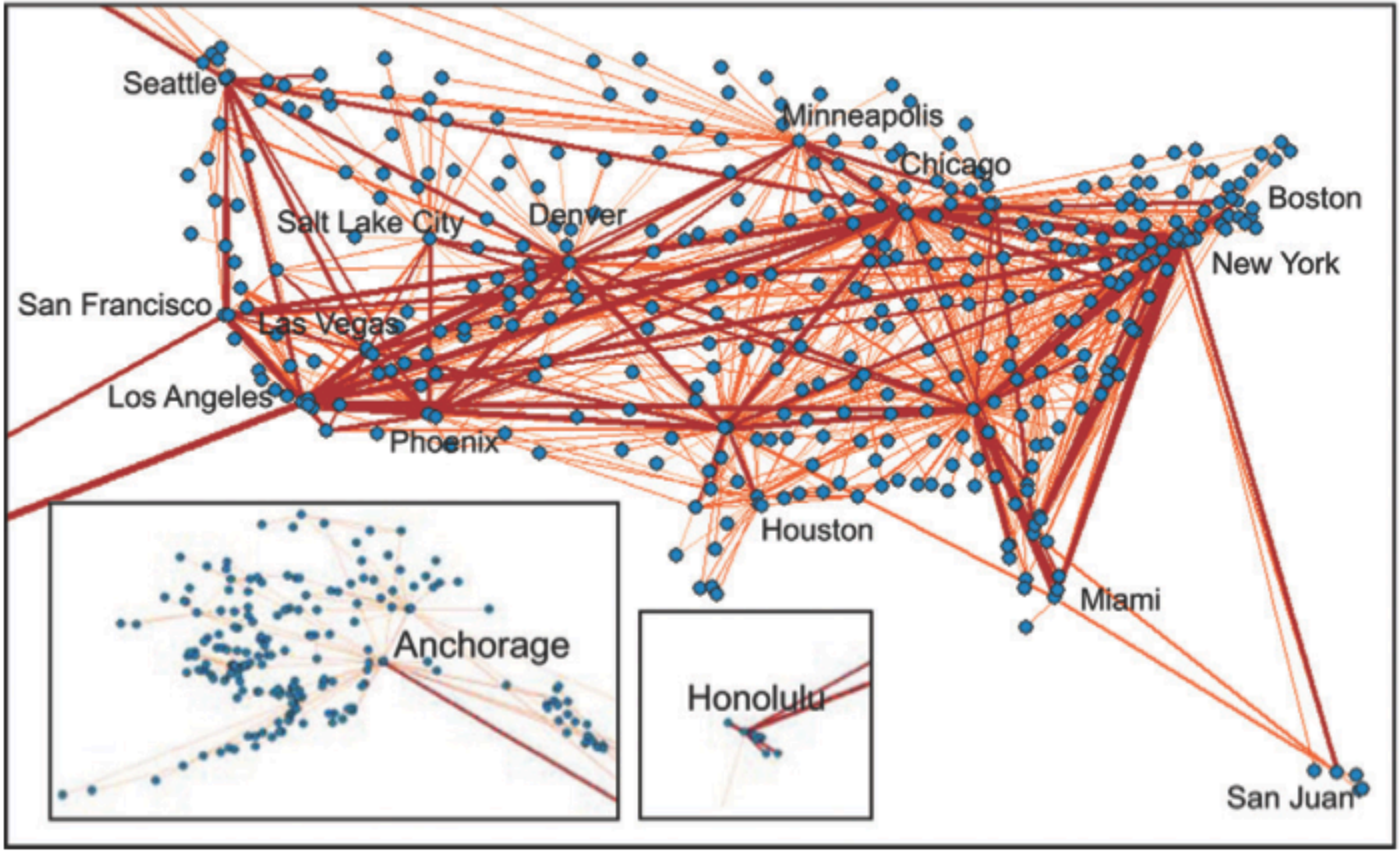


Fig. 4. Pajek representations (20) of disparity backbones. (*Left*) The $\alpha = 0.003$ multiscale backbone of the 2006 domestic segment of the U.S. airport transportation system. This disparity backbone includes entirely the top 10% of the heaviest edges. (*Right*) The $\alpha = 0.0008$ multiscale backbone of the Florida Bay ecosystem in the dry season. This disparity backbone includes entirely the top 40% of the heaviest edges. These disparity backbones correspond to points (b) for the U.S. airport network and (a) for the Florida Bay food web in Table 1 and Fig. 3. The connection with maximum weight for the U.S. airport network is Atlanta-Orlando, with value $\omega_{\max} = 1,290,488$ passengers/year and for the Florida Bay Food Web Free Bacteria to Water Flagellates with value $\omega_{\max} = 12.90$ mg C $y^{-1}m^{-2}$.

This approach was generalized in a paper by Radicchi et al that introduced the so-called Global Statistical Significance (GloSS) filter.

Radicchi, F., Ramasco, J.J. and Fortunato, S., 2011. Information filtering in complex weighted networks. Physical Review E, 83(4), p.046101.

Other recent approaches for link selection are

Dianati, N., 2016. Unwinding the hairball graph: pruning algorithms for weighted complex networks. Physical Review E, 93(1), p.012304.

Gemmetto, V., Cardillo, A. and Garlaschelli, D., 2017. Irreducible network backbones: unbiased graph filtering via maximum entropy. arXiv preprint arXiv:1706.00230.

Kobayashi, T., Takaguchi, T. and Barrat, A., 2018. The structured backbone of temporal social ties. arXiv preprint arXiv:1804.08828.

Marcaccioli, R. and Livan, G., 2018. A parametric approach to information filtering in complex networks: The Polya filter. arXiv preprint arXiv:1806.09893.

An approach investigating repeated events.

Quite appropriate for (repeated) events occurring in a temporal network.

This approach allows us to detect links that are statistically validated against a null hypothesis and therefore are elements of ***statistically validated networks***

Tumminello, M., Miccichè, S., Lillo, F., Piilo, J. and Mantegna, R.N., 2011. Statistically validated networks in bipartite complex systems. PloS one, 6(3), p.e17994.

Hatzopoulos, V., Iori, G., Mantegna, R.N., Miccichè, S. and Tumminello, M., 2015. Quantifying preferential trading in the e-MID interbank market. Quantitative Finance, 15(4), pp.693-710.

Li, M.X., Palchykov, V., Jiang, Z.Q., Kaski, K., Kertész, J., Miccichè, S., Tumminello, M., Zhou, W.X. and Mantegna, R.N., 2014. Statistically validated mobile communication networks: the evolution of motifs in European and Chinese data. New Journal of Physics, 16(8), p.083038.

Hypothesis testing is the process quantifying the evidence contained in observations, say $x_1, x_2, \ldots, x_n$, in support of a hypothesized theoretical model.

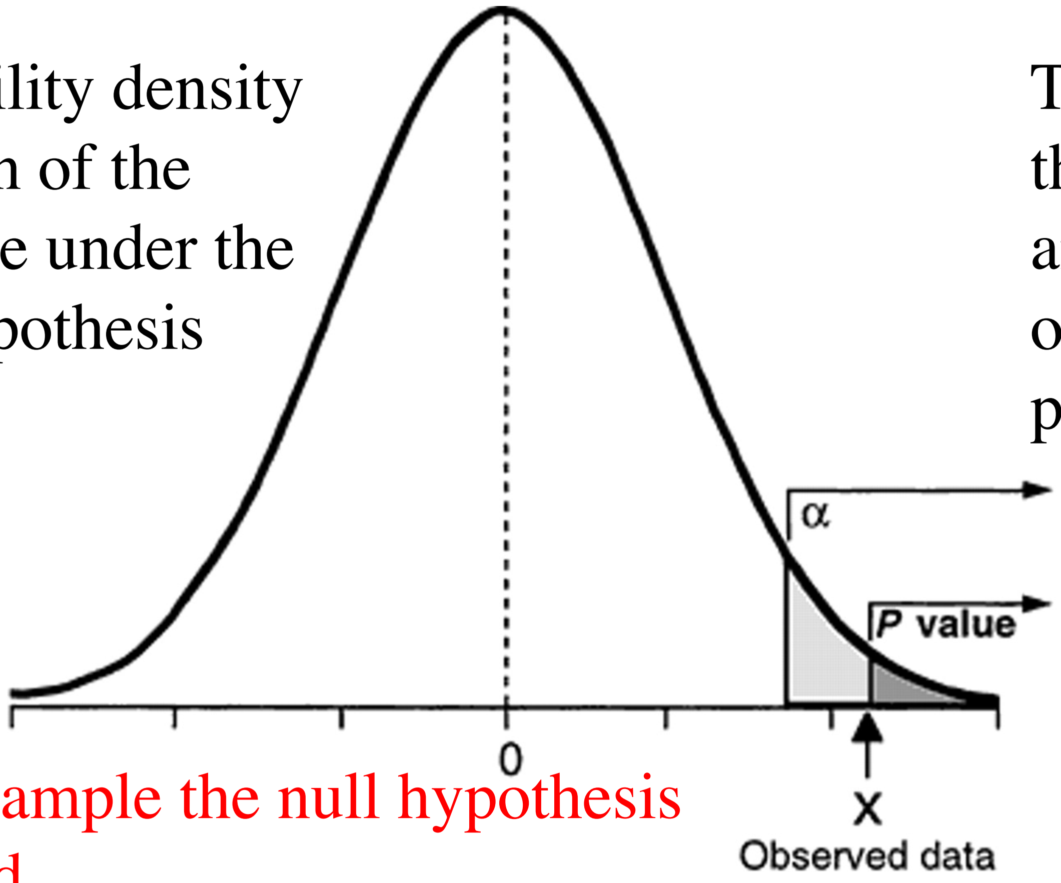
The theory is called the null hypothesis and it is typically indicated as H_0. Every other admissible hypothesis is called an alternative hypothesis.

Performing a hypothesis test generally requires computing a given test statistics $t_{\mathrm{obs}}=t(X)$ from the sample data and comparing it to a reference distribution F_{t} that describes the random behavior of the variable t in the case that H_0 were true.

If $F_t(t_{\mathrm{obs}})$ is sufficiently small (i.e. smaller than a previously defined statistical threshold) then the null hypothesis is rejected.

The *p*-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.

Probability density function of the outcome under the null hypothesis



The statistical threshold defines an interval of data on the theoretical pdf

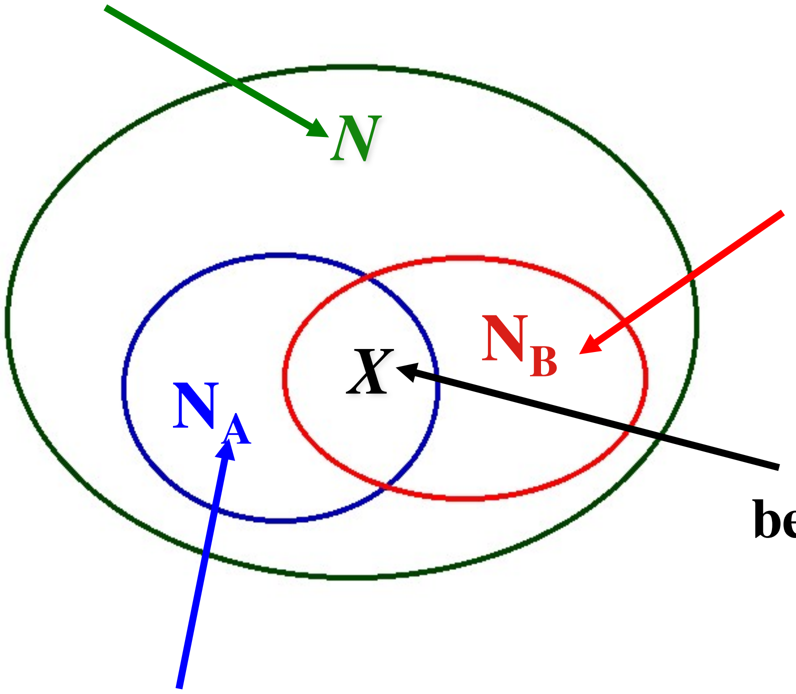
The gray area under the pdf gives the p-value associated with the observed data

In this example the null hypothesis is rejected

A statistical validation of a number of events observed between a pair of nodes

Suppose there are N events in the investigated set during the considered time frame. Suppose we are interested to evaluate the over-expression or under-expression of events occurring for each pair of nodes against a null hypothesis. For a given node, let us call N_A the number of events A has done and N_B the number of events B has done

Total # of events in the network



# of events N_B of node B

# of events between node A and node B

# of events N_A of node A

What is the probability of X under the null hypothesis of random matching?

The probability that X events are occurring between node A and node B is well approximated by the hypergeometric distribution

$$P(X \mid N, N_A, N_B) = \frac{\binom{N_A}{X}\binom{N - N_A}{N_B - X}}{\binom{N}{N_B}}$$

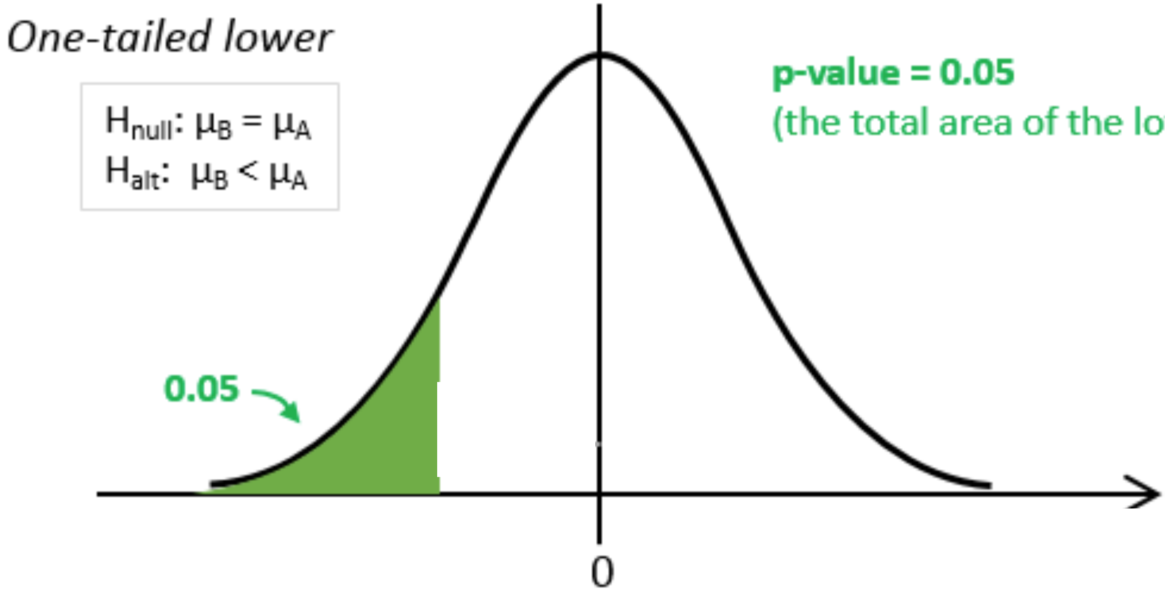
It is therefore possible to associate a p-value to an empirically observed value

p-value associated with a detection of co-occurrence ≥ X:

$$p = 1 - \sum_{i=0}^{X-1} \frac{\binom{N_A}{i}\binom{N - N_A}{N_B - i}}{\binom{N}{N_B}}$$

In addition to the detection of pairs of market members having over-expressed number of "HF trades" it is also possible to quantify the presence of *under-expressed* number of "HF trades".

p-value associated with a detection of co-occurrence $\leq$ X: $p = \sum_{i=0}^{X} \frac{\binom{N_A}{i}\binom{N-N_A}{N_B-i}}{\binom{N}{N_B}}$



p-value = 0.05
(the total area of the lower tip)

Familywise error correction

Analyzing a large network of N nodes with E edges requires E statistical tests (when we do a one tail test) and order N^2 tests (when we do a two tails test).

This number can be a quite large number and therefore to assess the statistical validity one needs to correct for familywise error.

Multiple comparison procedures require a multiple hypothesis test correction to avoid a large increase of false positive.

The most robust multiple hypothesis test correction is known as Bonferroni correction.

Bonferroni correction is performed as follows

1) let us call α the chosen statistical threshold of the single test;

2) let us call T the number of tests to be performed to obtain a given statistical conclusion;

3) the multiple hypothesis test correction is done by re-defining the statistical threshold for each test as $\alpha_B = \alpha/T$.

Heuristic motivation for the multiple hypothesis test correction.

Is a given investor advisor doing better that just providing random advice?

Let us consider the problem of predicting whether a given stock or index will rise (+) or fall (-) in 10 different events.

By advising randomly, what is the probability that the advisor is making at least 8 correct predictions out of the 10 cases?

The probability is

$$\frac{\binom{10}{10}+\binom{10}{9}+\binom{10}{8}}{2^{10}}=\frac{1+10+45}{1024}=0.0547$$

Let us consider now that we observe a panel of N=50 investment advisors. We assume that analyzing all of them and selecting the one with the best performance will provide an indication of robust deviation from the null hypothesis.

This line of reasoning is incorrect because does not consider that we are performing simultaneously a large number of tests and use implicitly a statistical threshold that was originally devised for a single test.

In fact, let us ask ourselves what is the probability that we observe at least one investor advisors making at least 8 correct predictions.

$q=1-0.0547$ is the probability of not making at least 8 correct predictions

$p=1-(1-0.0547)^{50}=0.9399$ is the probability that at least one investor advisor in the set of 50 is making at least 8 correct predictions.

An example of the Bonferroni correction.

Let us imagine that the 50 investor advisors present the following probability mass function of the number of correct predictions

# of correct predictions	0	1	2	3	4	5	6	7	8	9	10
# of investor advisors	0	0	1	3	7	12	15	8	2	1	1

Under the null hypothesis of random predictions the p-value associated with at least 10, 9, and 8 correct predictions are:

10 predictions: $p-value = \binom{10}{10} / 2^{10} = 0.000976$

9 predictions: $p-value = \left[\binom{10}{10} + \binom{10}{9}\right] / 2^{10} = 0.0107$

8 predictions: $p-value = \left[\binom{10}{10} + \binom{10}{9} + \binom{10}{8}\right] + /2^{10} = 0.0547$

By assuming a single test threshold θ=0.05 we would reject the case of 9 and 10 predictions

But with the Bonferroni correction the threshold to consider is $\alpha_B=0.05/50=0.001$ and the null hypothesis is only rejected for the case of 10 corrected predictions

Problems related with the control of the familywise error rate. Let us consider the general problem of classifying the rejection of a null hypothesis done by estimating the so-called confusion matrix.

The confusion matrix contains information about

f_{11} or TP # of objects that are of class 1 and are predicted in class 1

f_{10} or FN # of objects that are of class 1 and are predicted in class 0

f_{01} or FP # of objects that are of class 0 and are predicted in class 1

f_{00} or TN # of objects that are of class 0 and are predicted in class 0

		Predicted	
		Class 1	Class 0
Actual	Class 1	f_{11}	f_{10}
	Class 0	f_{01}	f_{00}

		Predicted	
		Class 1	Class 0
Actual	Class 1	TP	FN
	Class 0	FP	TN

From the confusion matrix a series of indicators are usually extracted.

The most common of them are:

$$\text{Accuracy} = \frac{\text{\# of correct predictions}}{\text{total predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}} = \frac{TP + TN}{TP + FN + FP + TN}$$

$$\text{Error rate} = \frac{\text{\# of wrong predictions}}{\text{total predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}} = \frac{FN + FP}{TP + FN + FP + TN}$$

$$\text{Sensitivity} = \frac{f_{11}}{f_{11} + f_{10}} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{f_{00}}{f_{00} + f_{01}} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{f_{11}}{f_{11} + f_{01}} = \frac{TP}{TP + FP}$$

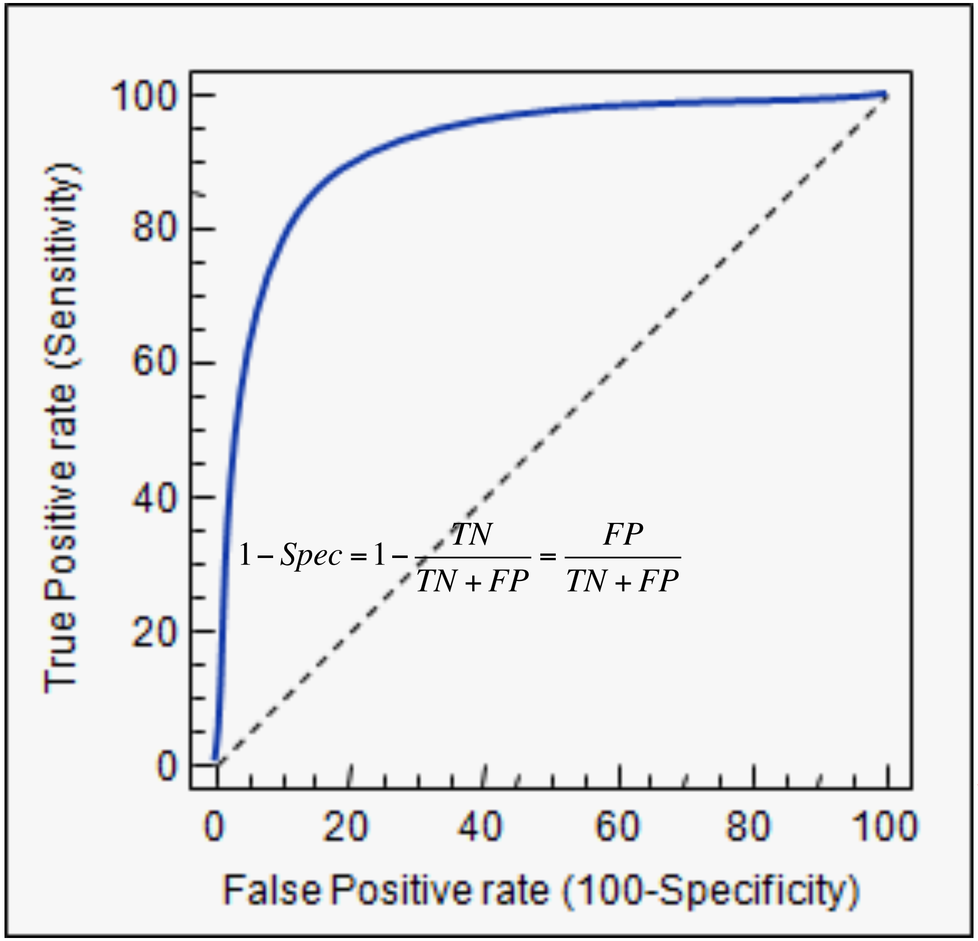
Note that sensitivity (also called recall) and specificity are difficult to maximize simultaneously. By looking for approaches increasing sensitivity, for example increasing the number of TP, as a result one also typically increases FP therefore decreasing specificity.

The Receiver Operating Characteristic (ROC) curve

True Positive Rate (Sensitivity) (Recall)

$$\frac{TP}{TP+FN}=\frac{f_{11}}{f_{11}+f_{10}}$$

fraction of positive examples predicted correctly by the model



$$1-Spec=1-\frac{TN}{TN+FP}=\frac{FP}{TN+FP}$$

		P	
		Class 1	Class 0
A	Class 1	f_{11}	f_{10}
	Class 0	f_{01}	f_{00}

False Positive Rate (1-Specificity)

$$\frac{FP}{TN+FP}=\frac{f_{01}}{f_{00}+f_{01}}$$

fraction of negative examples incorrectly predicted as positive by the model

Bonferroni’s correction is minimizing the number of False Positive. It usually provide a very high precision and low False positive rate. However, this is typically done at the cost of a large number of False Negative which is seriously decreasing the sensitivity and accuracy of the test.

Several methods have therefore been proposed to increase the sensitivity and accuracy of the procedure of controlling the familywise error rate without affecting significantly the False Positive rate. Currently the more widely used method is the one controlling the false discovery rate.

Benjamini, Yoav, and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal Statistical Society. Series B (Methodological) (1995): 289-300.

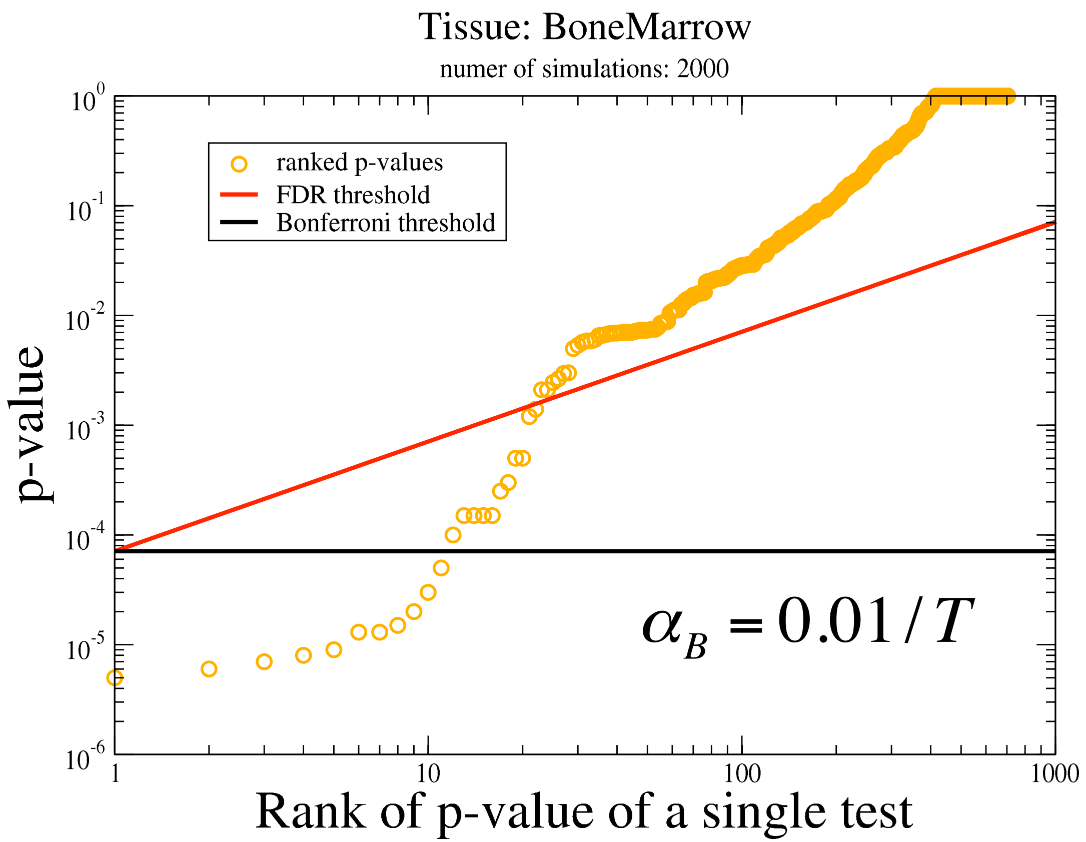
An example of the Benjamini-Hochberg procedure of controlling the False Discovery Rate.

0) determine the Bonferroni threshold α_B;

1) rank the p-values of the T tests performed;

2) consider the function $\alpha_k = k\, \alpha_B$ and select the k smallest integer such the function intersects the ranked p-values

3) consider rejected all test with p-value less than α_k^* crossing value of the two curves.

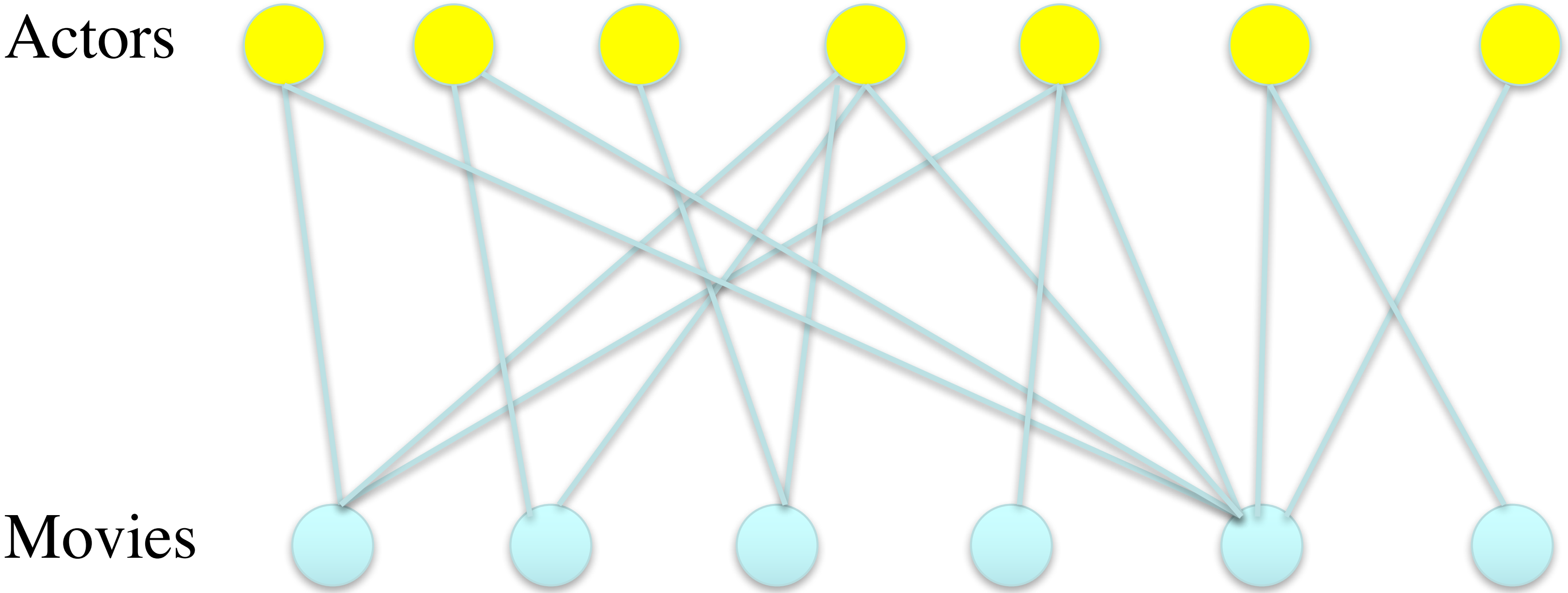


Filtering of a large bipartite system

Example: The dataset of Movies and Actors

All world movies present in the IMDB database and produced during the period 1990-2008.

The dataset comprises 89605 movies realized in 158 countries. 412,143 actors have played in these movies.



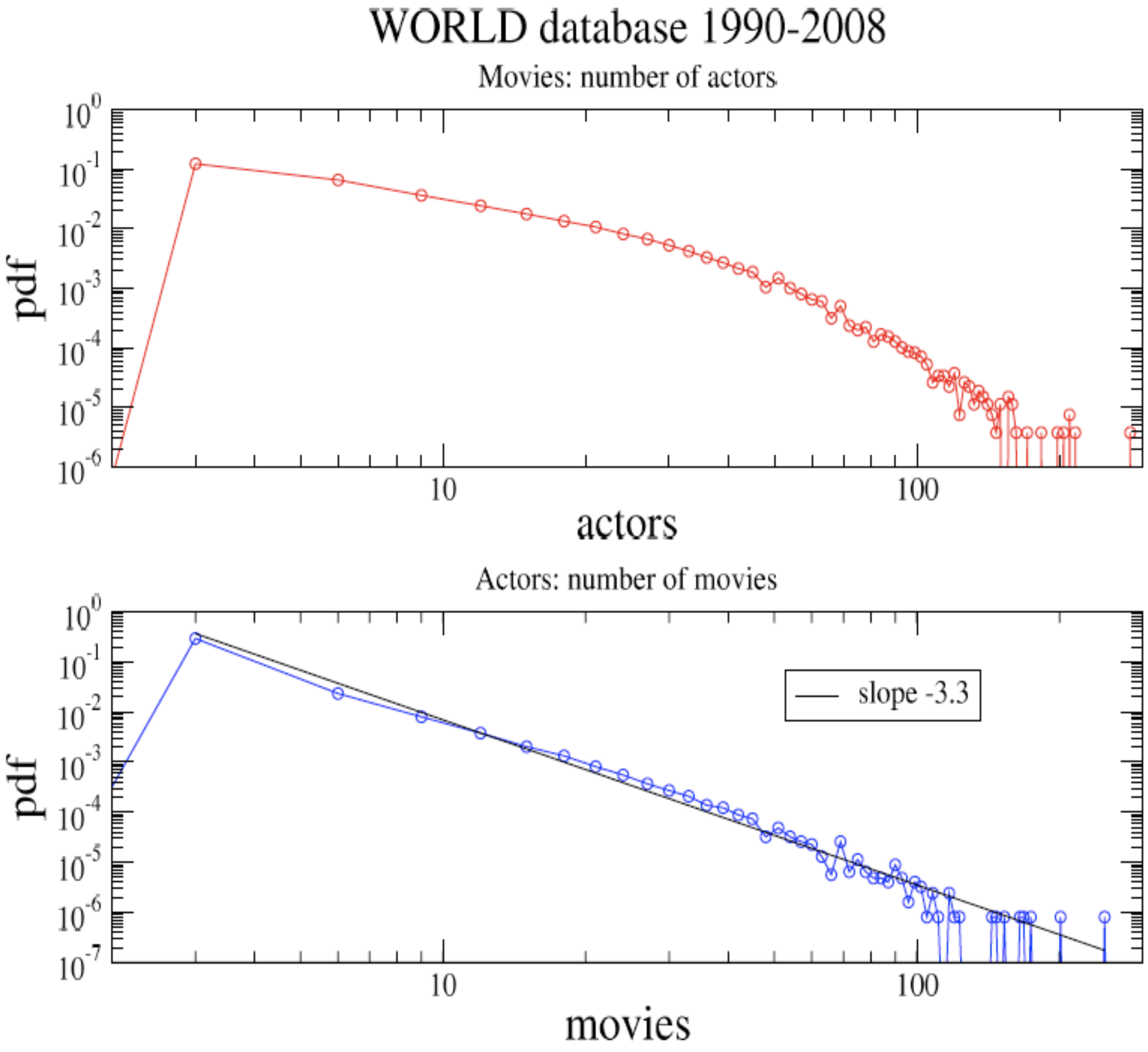
Actors

Movies

A high degree of heterogeneity is present in this dataset

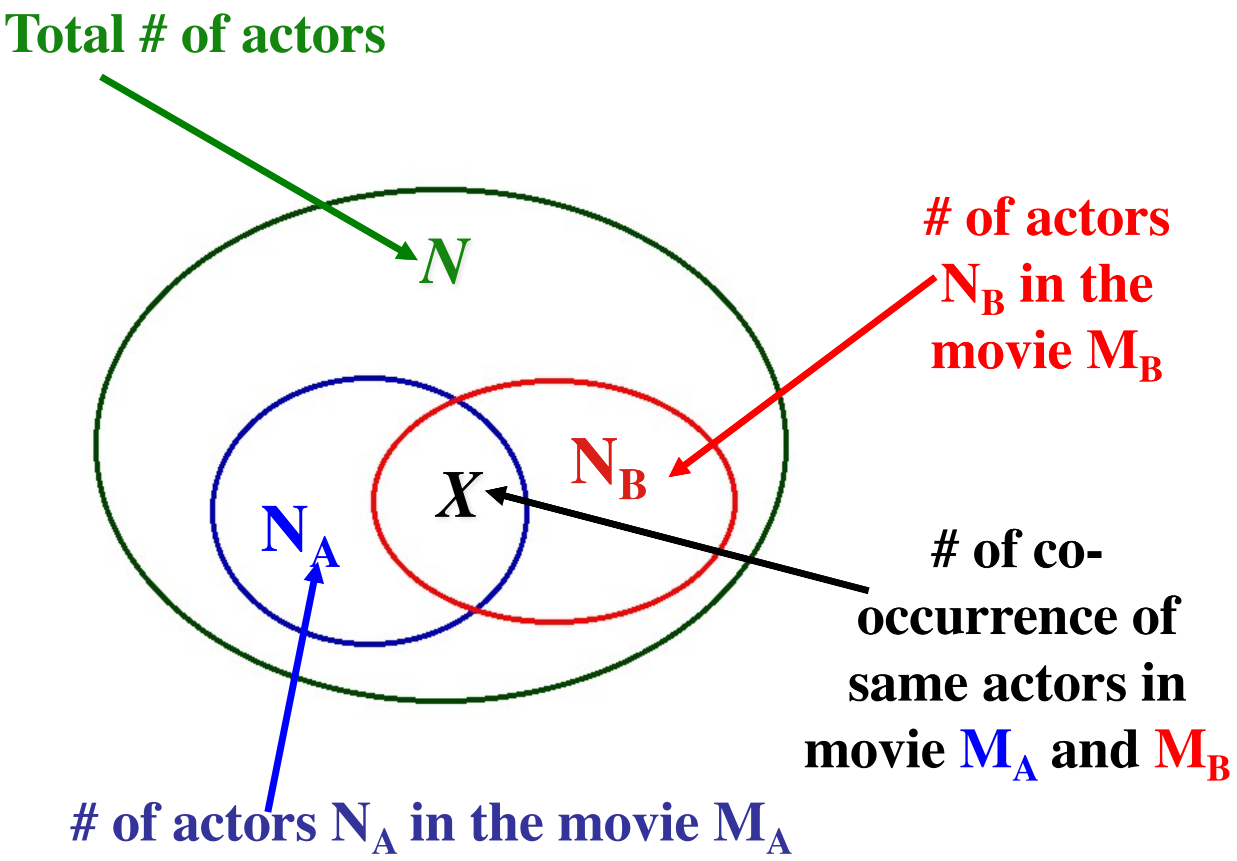
Authors investigated the projected network of movies

Tumminello M, Miccichè S, Lillo F, Piilo J, Mantegna RN (2011) Statistically Validated Networks in Bipartite Complex Systems. PLoS ONE 6(3): e17994.



A statistical validation in bipartite networks

Suppose there are N actors in the investigated set. Suppose we are interested to evaluate co-occurrence against a null hypothesis of random selection of actors playing both in movie $\mathbf{M_A}$ and $\mathbf{M_B}$. $\mathbf{M_A}$ has a cast of $\mathbf{N_A}$ actors whereas $\mathbf{M_B}$ has a cast of $\mathbf{N_B}$ actors. Let us call X the co-occurrence of the same **X** actors in movie $\mathbf{M_A}$ and $\mathbf{M_B}$.


Total # of actors
N
# of actors N_B in the movie M_B
N_B
X
N_A
# of co-occurrence of same actors in movie M_A and M_B
# of actors N_A in the movie M_A

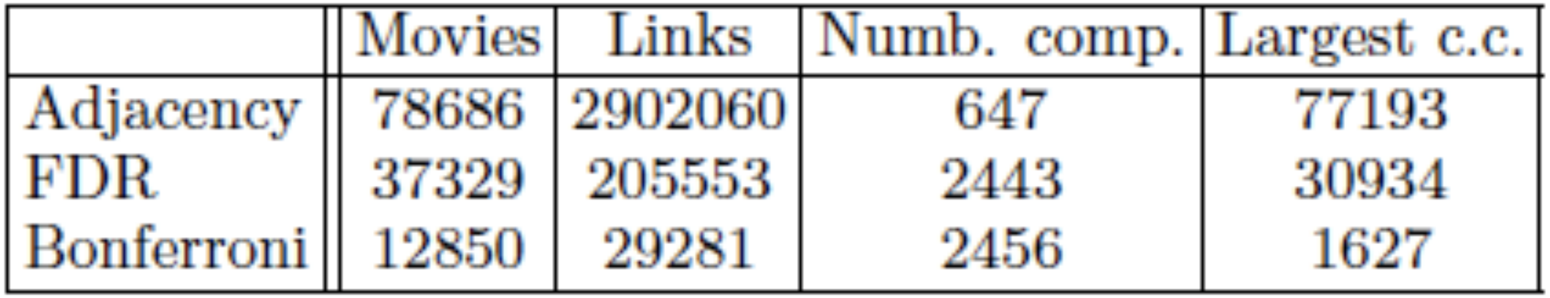
We address the **statistically validated network** obtained with the Bonferroni correction as the **Bonferroni network**

We address the **statistically validated network** obtained by controlling the FDR as the **FDR network**

Original, FDR and Bonferroni Network

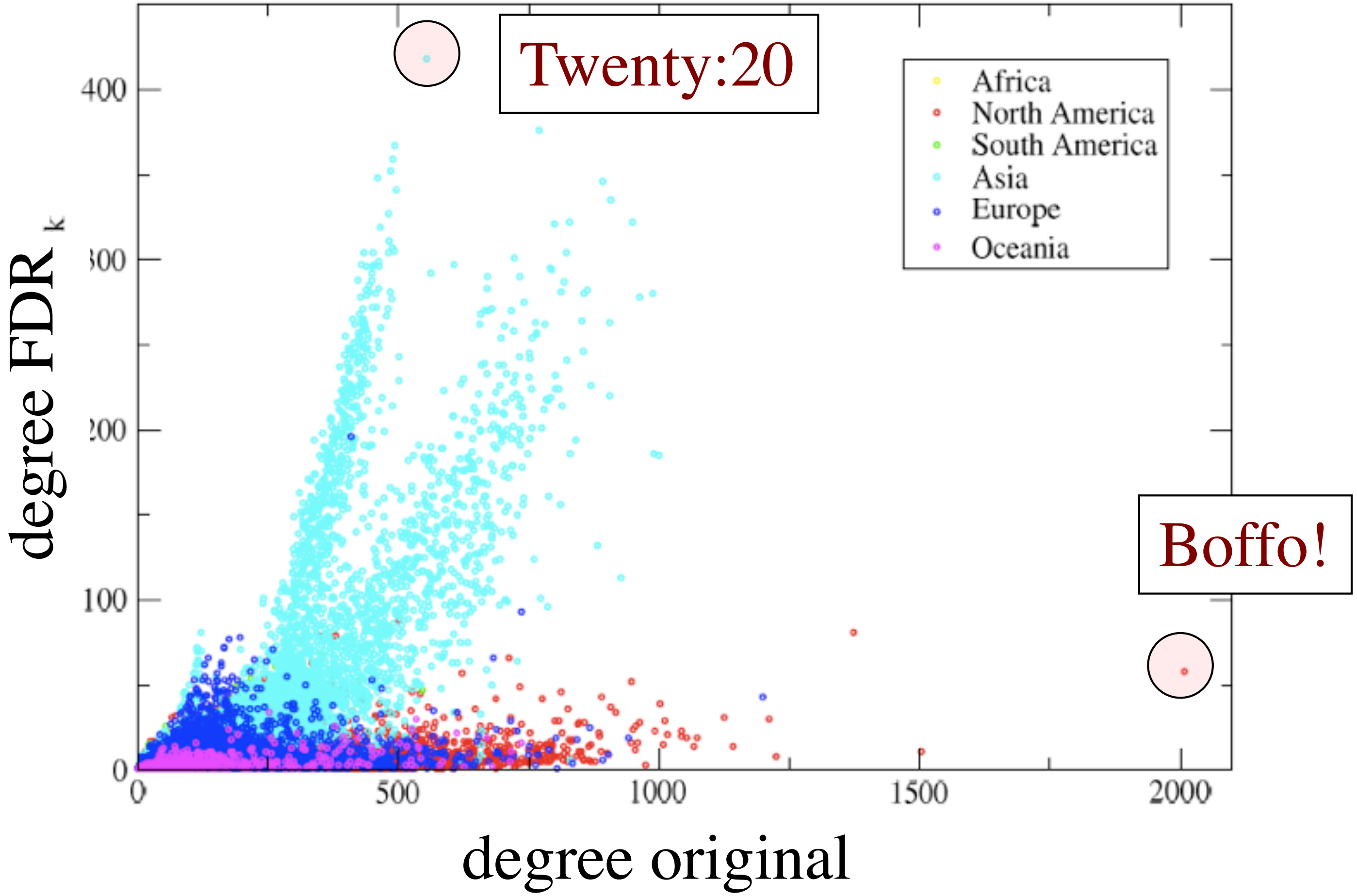
In the adjacency network two movies are connected when at least one actor is playing in both movies.

In the FDR and Bonferroni networks only a fraction of these links are statistically validated by using the Hypergeometric distribution and the multiple test correction

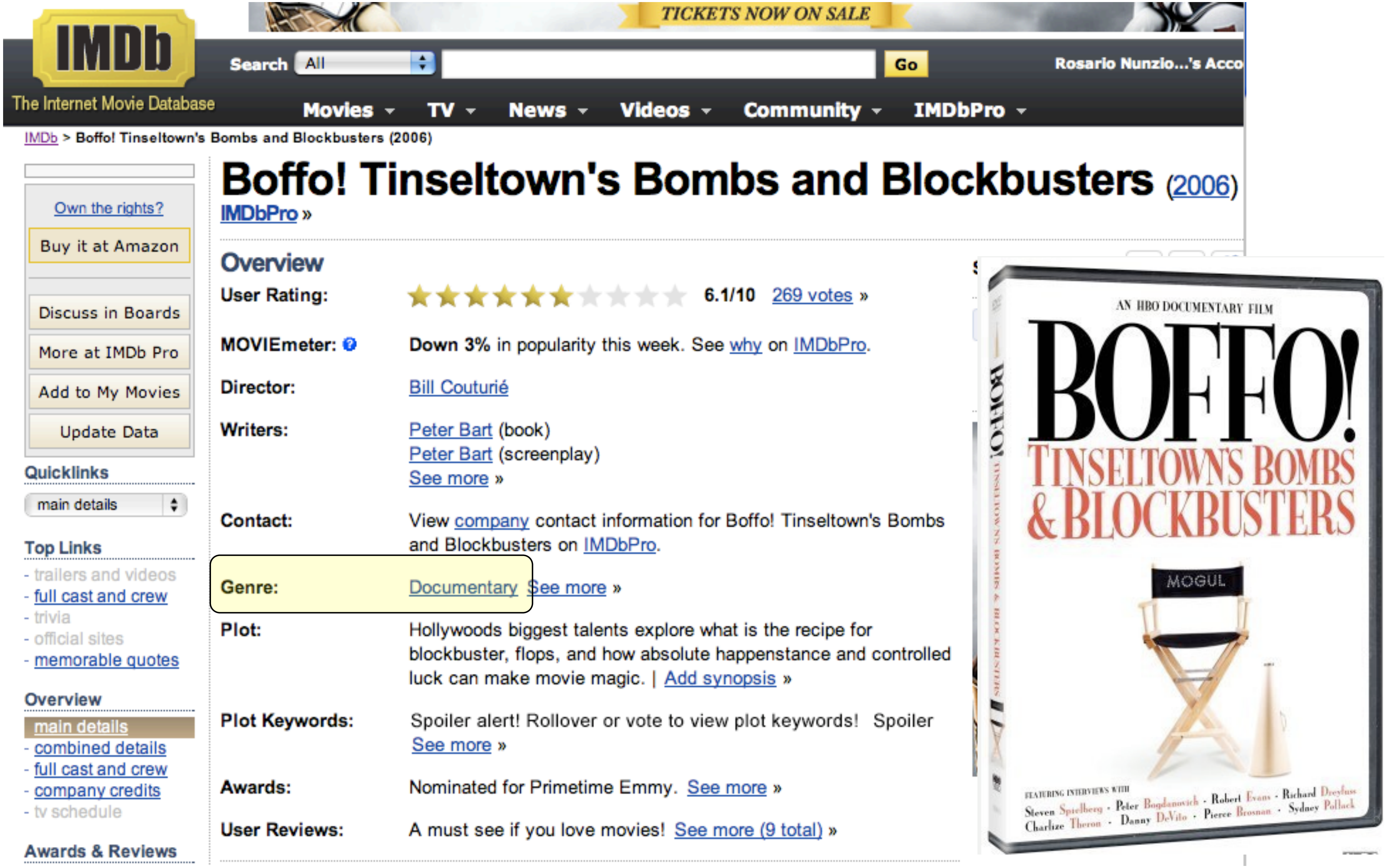


	Movies	Links	Numb. comp.	Largest c.c.
Adjacency	78686	2902060	647	77193
FDR	37329	205553	2443	30934
Bonferroni	12850	29281	2456	1627

The degree profile of movies shows the existence of regions of the original network with different topology



The movie with highest degree in the adjacency network (134 actors): Degree 2008



IMDb > Boffo! Tinseltown's Bombs and Blockbusters (2006)

Boffo! Tinseltown's Bombs and Blockbusters (2006)

IMDbPro »

Overview

User Rating: ★★★★★★☆☆☆☆ 6.1/10 269 votes »

MOVIEmeter: Down 3% in popularity this week. See why on IMDbPro.

Director: Bill Couturié

Writers: Peter Bart (book)
Peter Bart (screenplay)
See more »

Contact: View company contact information for Boffo! Tinseltown's Bombs and Blockbusters on IMDbPro.

Genre: Documentary See more »

Plot: Hollywoods biggest talents explore what is the recipe for blockbuster, flops, and how absolute happenstance and controlled luck can make movie magic. | Add synopsis »

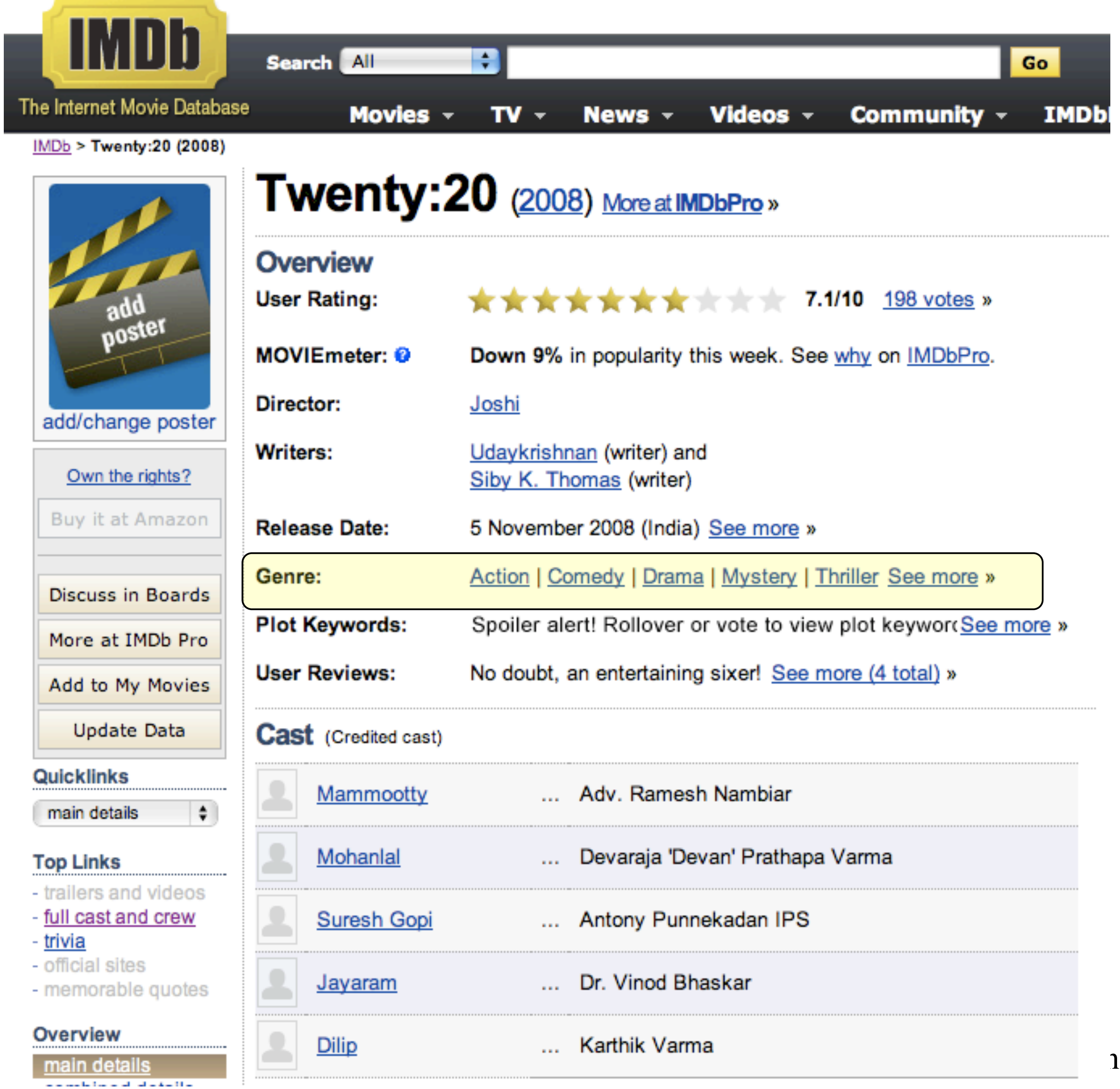
Plot Keywords: Spoiler alert! Rollover or vote to view plot keywords! Spoiler
See more »

Awards: Nominated for Primetime Emmy. See more »

User Reviews: A must see if you love movies! See more (9 total) »

AN HBO DOCUMENTARY FILM
BOFFO!
TINSELTOWN'S BOMBS & BLOCKBUSTERS
MOGUL
FEATURING INTERVIEWS WITH
Steven Spielberg · Peter Bogdanovich · Robert Evans · Richard Dreyfuss
Charlize Theron · Danny DeVito · Pierce Brosnan · Sydney Pollack

The movie with highest degree in the FDR and Bonferroni network (56 actors)



IMDb > Twenty:20 (2008)

Twenty:20 (2008) More at IMDbPro »

Overview

User Rating:	★★★★★★★☆☆☆ 7.1/10 198 votes »
MOVIEmeter:	Down 9% in popularity this week. See why on IMDbPro.
Director:	Joshi
Writers:	Udaykrishnan (writer) and Siby K. Thomas (writer)
Release Date:	5 November 2008 (India) See more »
Genre:	Action | Comedy | Drama | Mystery | Thriller See more »
Plot Keywords:	Spoiler alert! Rollover or vote to view plot keywor(See more »
User Reviews:	No doubt, an entertaining sixer! See more (4 total) »

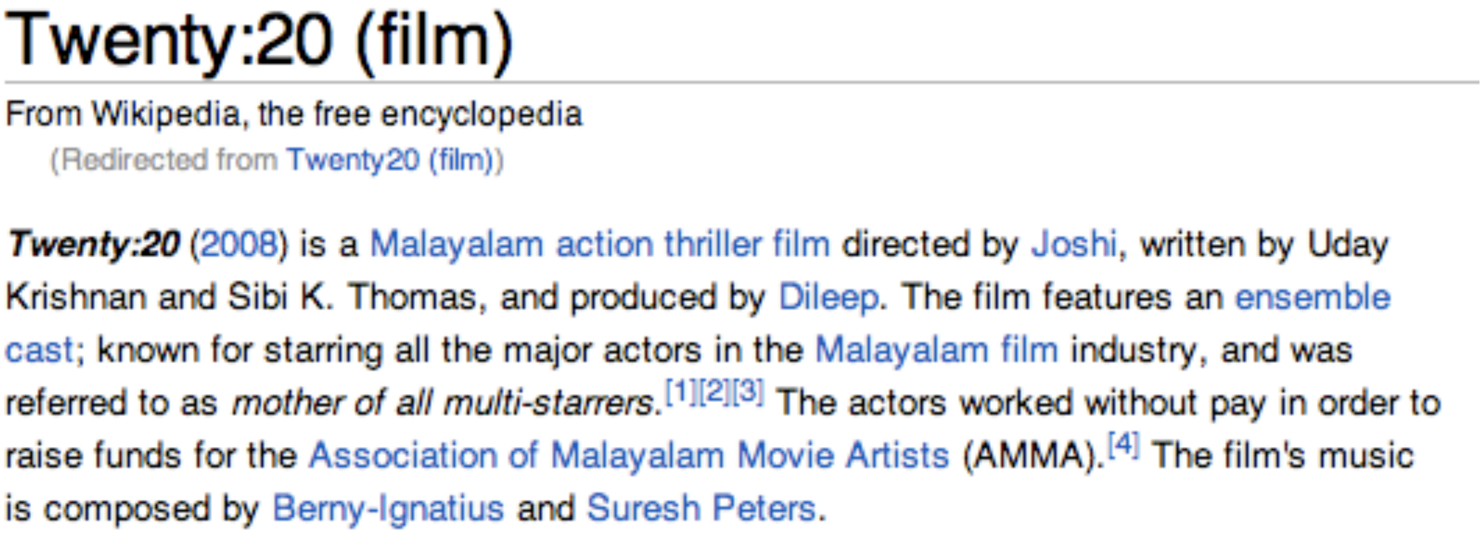
Cast (Credited cast)

Mammootty	...	Adv. Ramesh Nambiar
Mohanlal	...	Devaraja 'Devan' Prathapa Varma
Suresh Gopi	...	Antony Punnekadan IPS
Jayaram	...	Dr. Vinod Bhaskar
Dilip	...	Karthik Varma

Degree FDR 418

Degree Bonferroni 254

Twenty:20

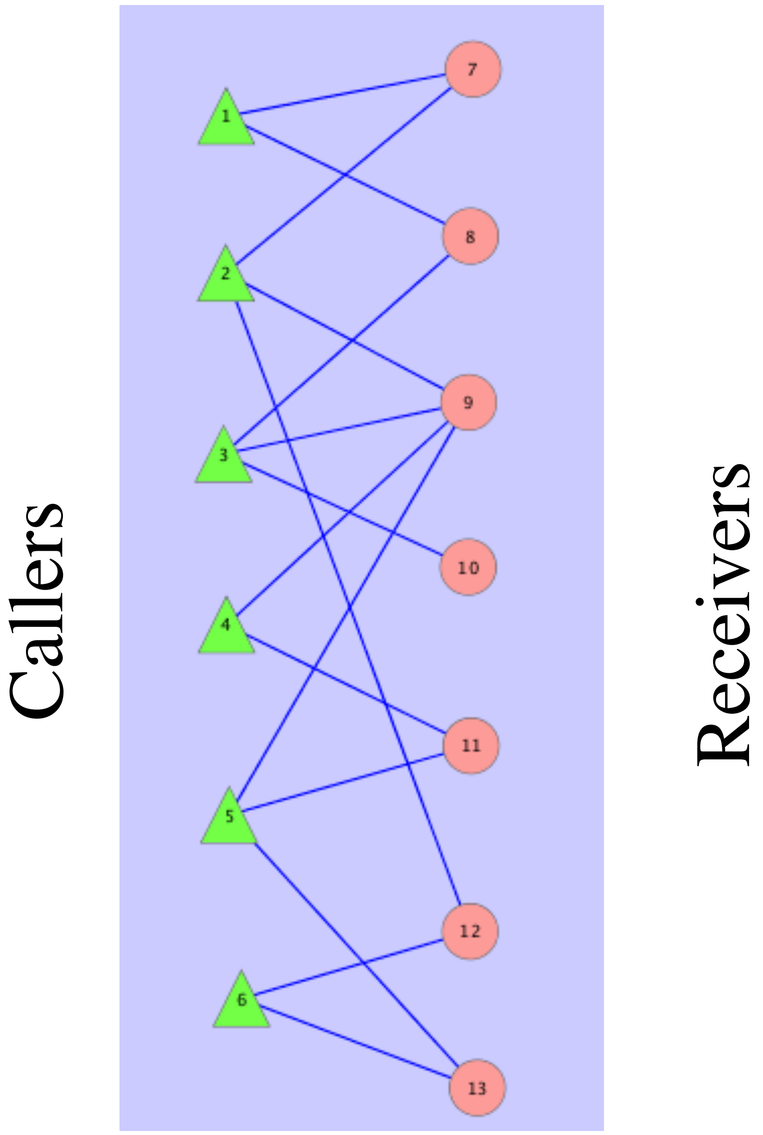


Twenty:20 (film)

From Wikipedia, the free encyclopedia
(Redirected from Twenty20 (film))

Twenty:20 (2008) is a Malayalam action thriller film directed by Joshi, written by Uday Krishnan and Sibi K. Thomas, and produced by Dileep. The film features an ensemble cast; known for starring all the major actors in the Malayalam film industry, and was referred to as *mother of all multi-starrers.*[1][2][3] The actors worked without pay in order to raise funds for the Association of Malayalam Movie Artists (AMMA).[4] The film's music is composed by Berny-Ignatius and Suresh Peters.

Network filtering in directed networks: mobile phone communications



Within a selected time window (daily, weekly, monthly, complete) we build a bipartite network of the calls occurring between mobile phone subscribers.

One key point is pre-processing the bipartite network to select calls that are originating from the underlying social network.

In previous studies the pre-processing was often performed by only considering reciprocated calls.

Filtering a mobile phone communications network

In the study of "social relationships, private communication is needed; however, experience tells us that sometimes phones registered as private are used for professional purposes, like in call centers or marketing and information campaigns. In fact, the presence of large spurious communication hubs, e.g. large call centers, significantly alters the statistics of 3-motifs (and, more generally, of any class of motifs). Dialing wrong numbers is another possible source of false links. In addition, the usual corruption arising during coding, transferring and processing data can also take place. Unless data are cleaned, spurious links could be misinterpreted as real social relationships. This problem is part of the general topic of information filtering in complex networks with strong inhomogeneities".

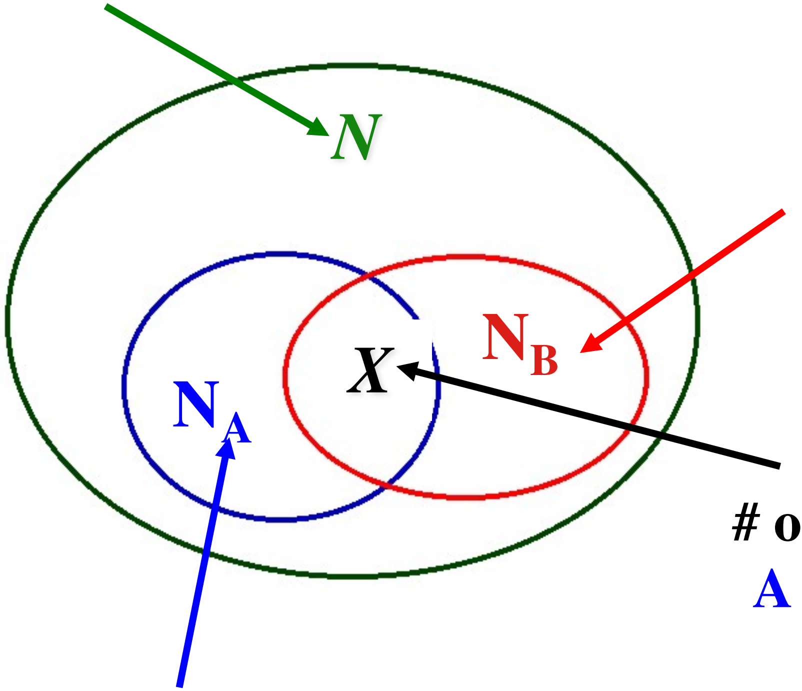
Ming-Xia Li, V. Palchykov, Z.-Q. Jiang, K. Kaski, J. Kertész, S. Miccichè, M. Tumminello, W.-X. Zhou and R.N. Mantegna, Statistically validated mobile communication networks: the evolution of motifs in European and Chinese data, New Journal of Physics 16 (2014) 083038

Ming-Xia Li, Z.-Q. Jiang, W.-J. Xie, S. Miccichè, M. Tumminello, W.-X. Zhou & R. N. Mantegna, A comparative analysis of the statistical properties of large mobile phone calling networks, Scientific Reports 4, Article number: 5132 (2014)

A statistical validation in directed bipartite networks

Suppose there are **N** calls in the investigated set. Suppose we are interested to evaluate the over-expression of a number of calls occurring between each pair of subscribers (when **A** is caller and **B** receiver) against a null hypothesis taking into account their calling heterogeneity. For a given subscriber, let us call $\mathbf{N_A}$ the number of calls **A** does as a caller and $\mathbf{N_B}$ the number of calls **B** does as a receiver

Total # of calls



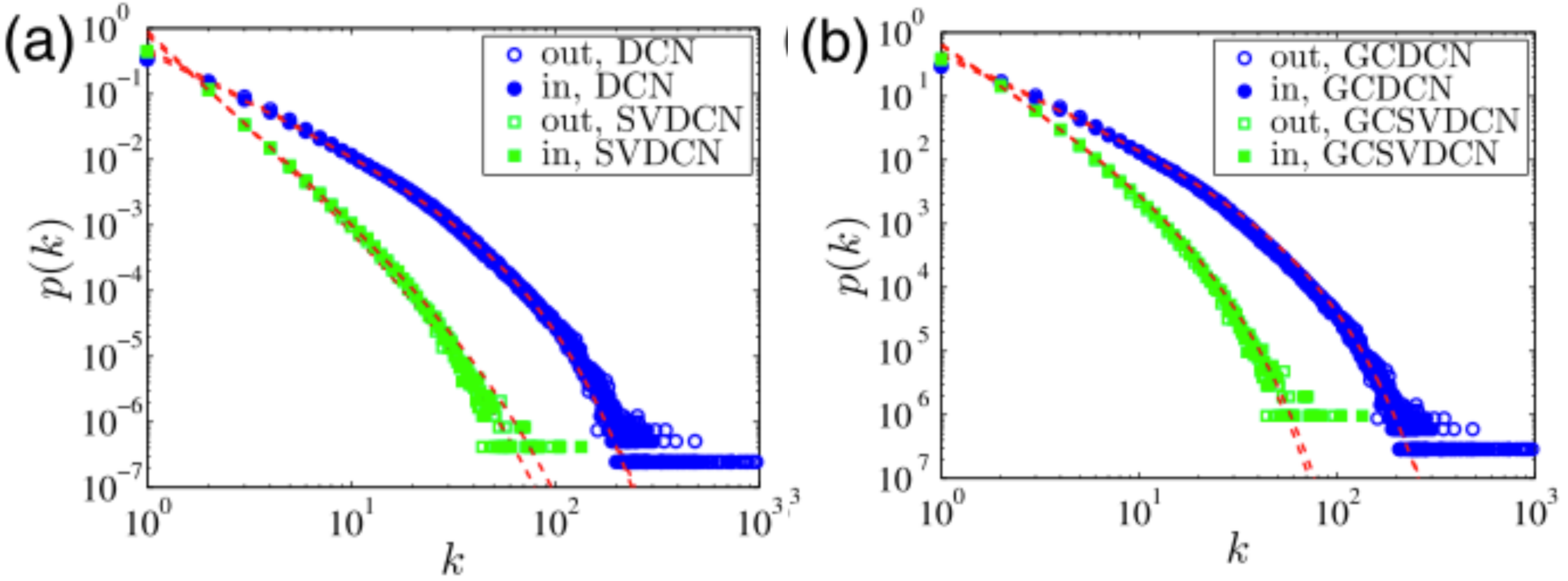
# of calls N_B made by subscriber B as a receiver

# of calls between A (caller) and B (receiver)

# of calls N_A made by subscriber A (caller)

Degree dist. in original and statistically validated networks

DCN: Directed Calling Network SV Statistically Validated



GC Giant Component

Chinese mobile phone operator

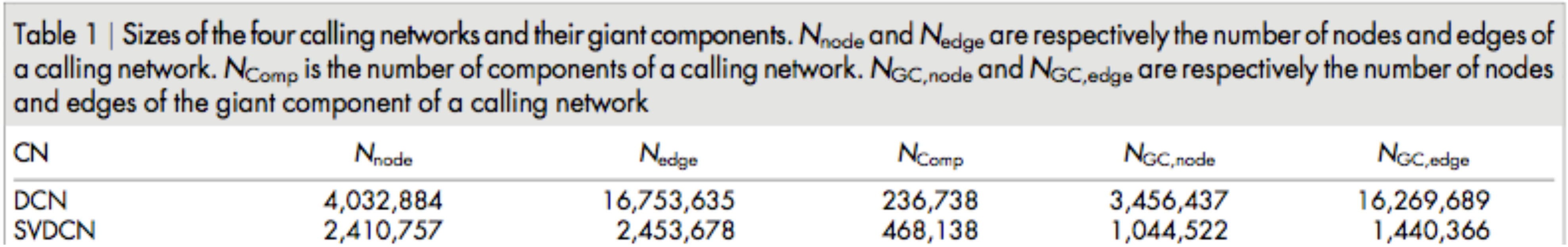
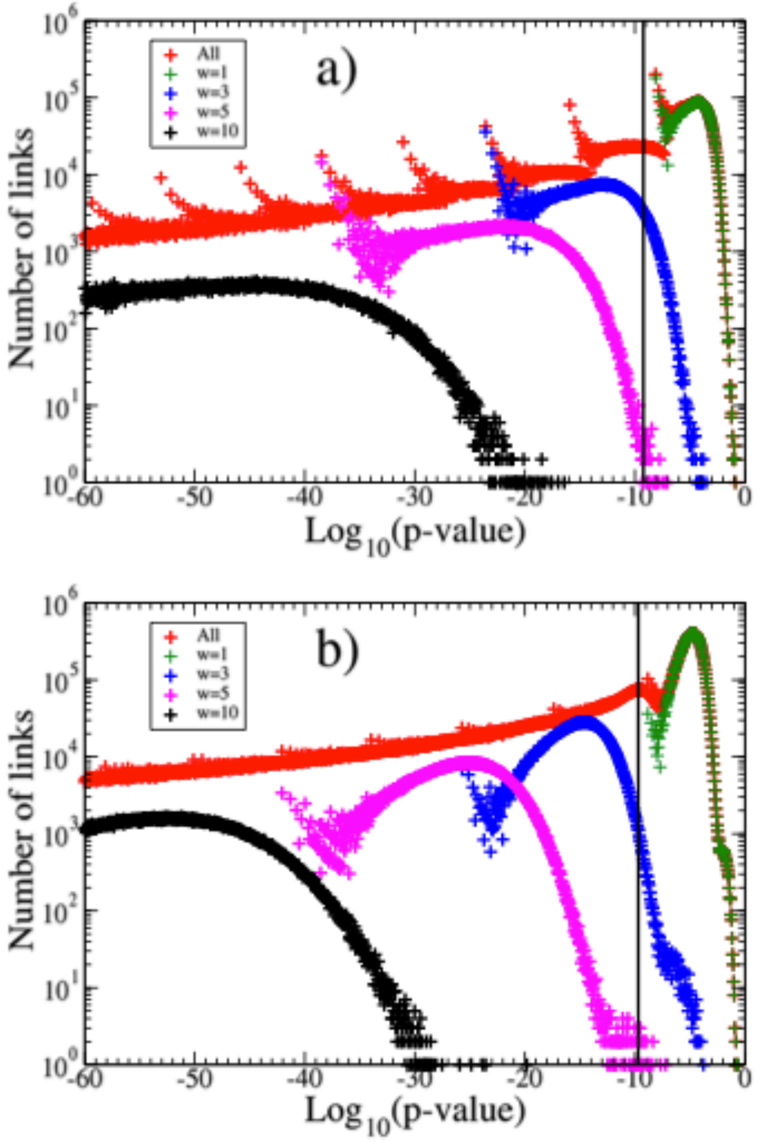


Table 1 | Sizes of the four calling networks and their giant components. N_{node} and N_{edge} are respectively the number of nodes and edges of a calling network. N_{Comp} is the number of components of a calling network. $N_{GC,node}$ and $N_{GC,edge}$ are respectively the number of nodes and edges of the giant component of a calling network

CN	N_{node}	N_{edge}	N_{Comp}	$N_{GC,node}$	$N_{GC,edge}$
DCN	4,032,884	16,753,635	236,738	3,456,437	16,269,689
SVDCN	2,410,757	2,453,678	468,138	1,044,522	1,440,366



Chinese mobile phone operator

European mobile phone operator

Figure 1. Number of links as a function of the *p*-value for the Chinese (panel (a)) and European (panel (b)) datasets. The red symbols describe the histogram for all links. Symbols of different color refer to the number of links of pairs of callers and receivers with weight equal to 1 (green), 3 (blue), 5 (purple) and 10 (black). The vertical line indicates the Bonferroni threshold. Links located to the left of the threshold are retained in the Bonferroni network. The network is obtained by considering the entire period. Only links between subscribers are considered.

The Bonferroni correction is guaranteeing high precision P=TP/(TP+FP) (i.e. a very low value of FP)

but this can be provided at a cost of low accuracy

Acc=(TP+TN)/(TP+TN+FP+FN)

due to a large number of FN.

However, in spite of this limitation the Bonferroni networks can be highly informative with respect to specific scientific questions when a high degree of precision is needed to obtain accurate results.

Effect of filtering on the measurement of some social sensitive network metric

Because in social networks and in computational social sciences there is a strong interest in the detection and interpretation of triads (3-motifs), i.e. mesoscopic structure of the networks

The set of isomorphic triads (3-motifs) is

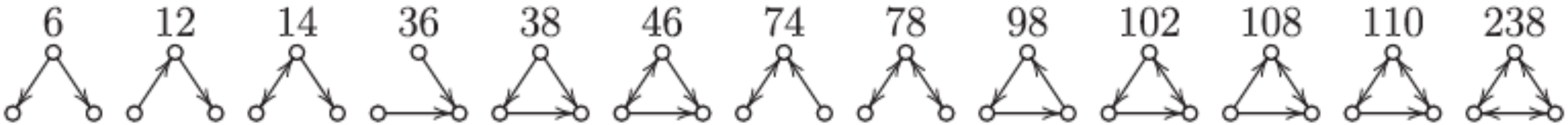
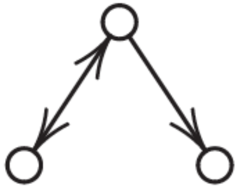
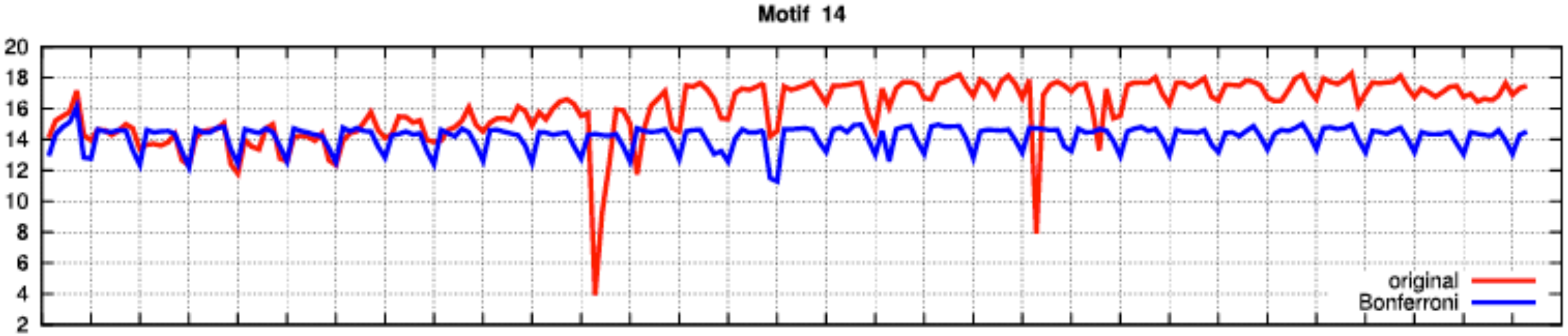


Figure 4. List of directed 3-motifs.

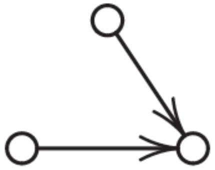
14

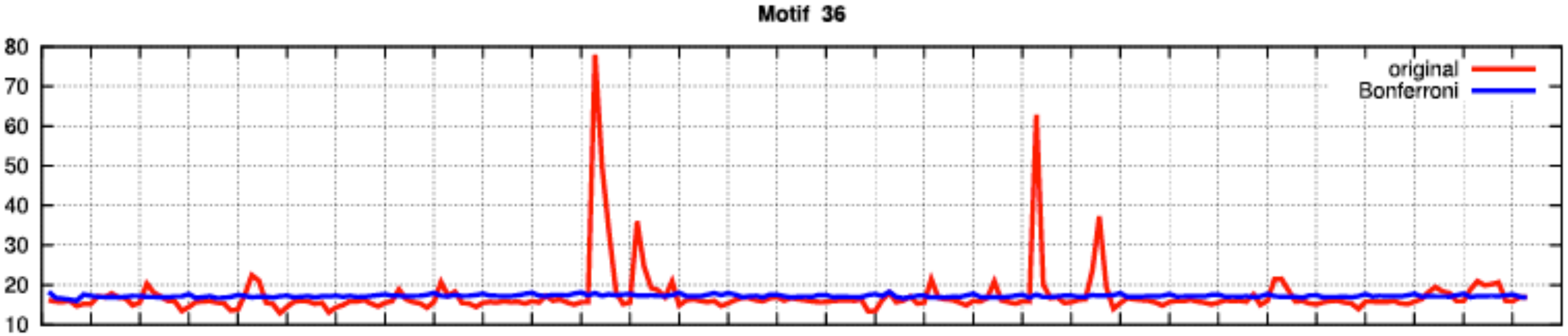


%

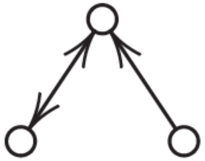


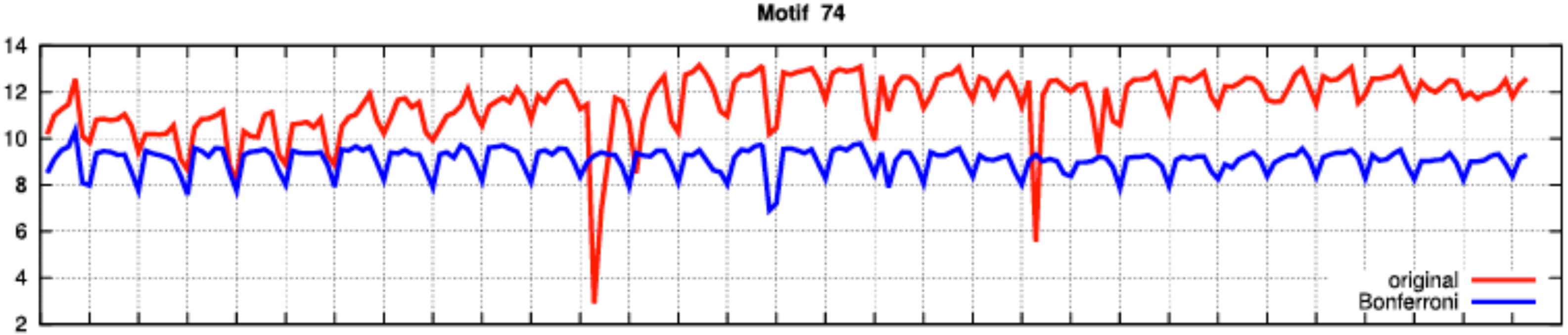
36



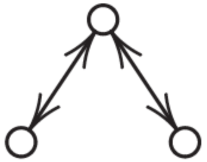


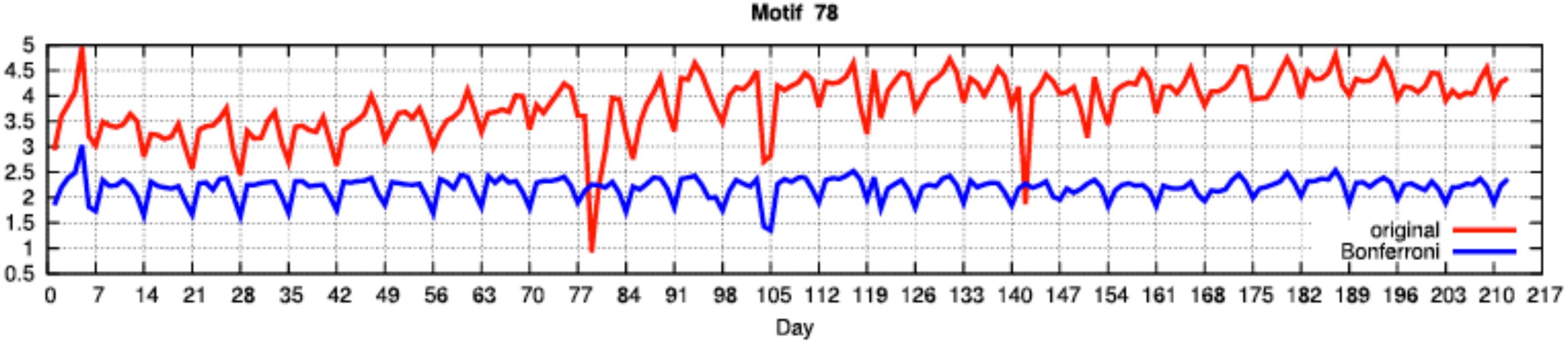
74





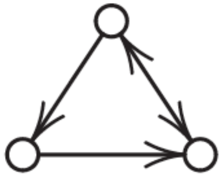
78

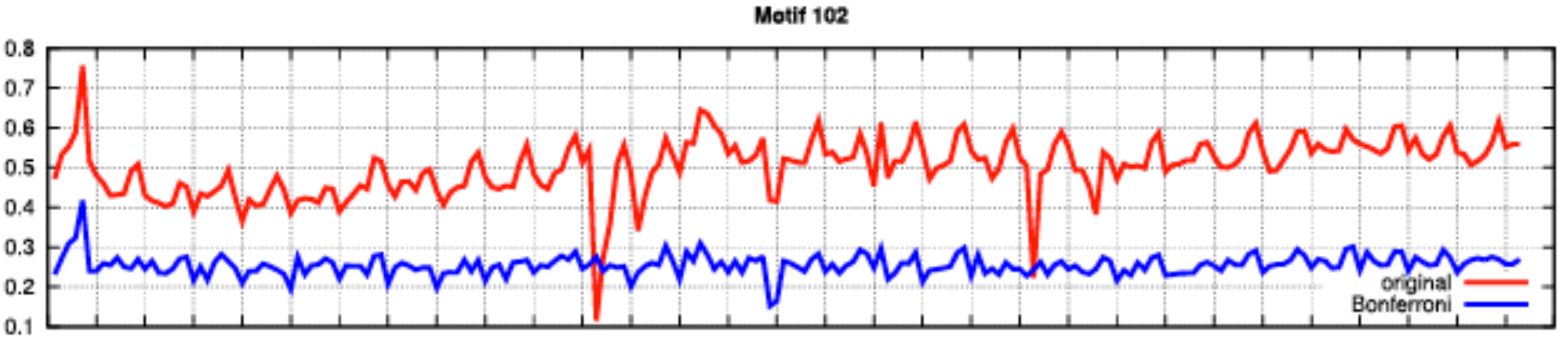




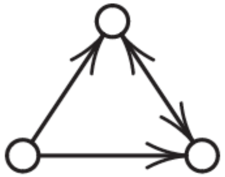
21 July, 2018

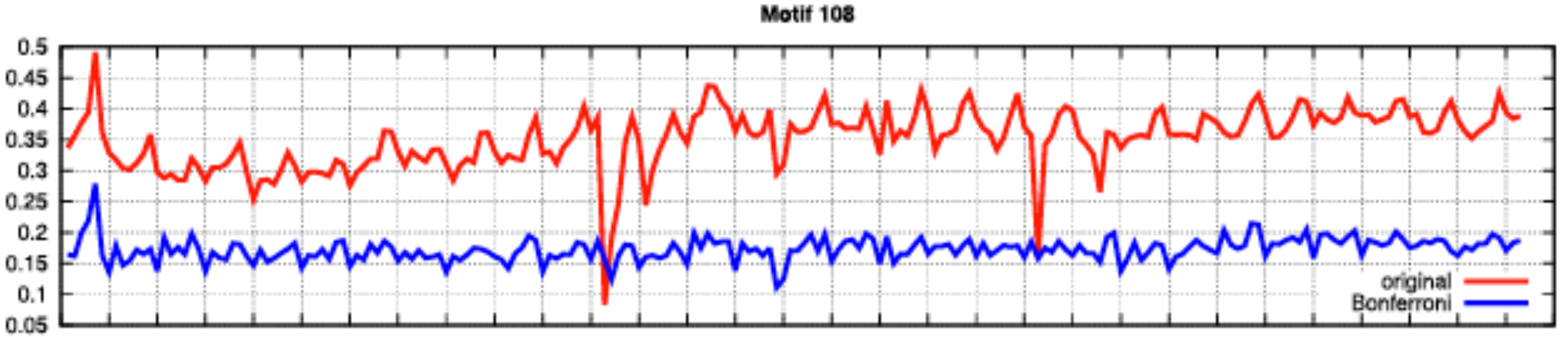
102 %



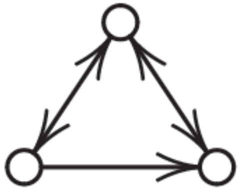

Motif 102

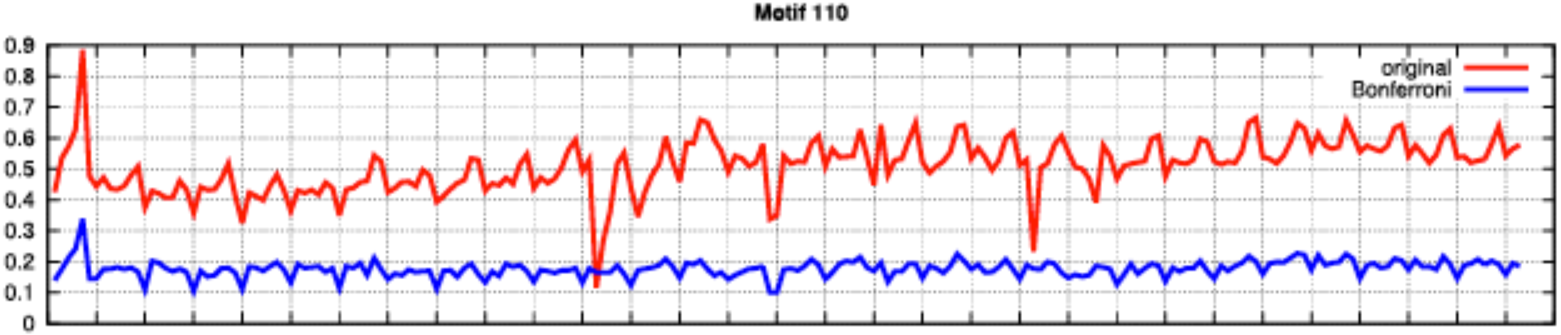
108



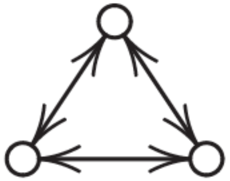

Motif 108

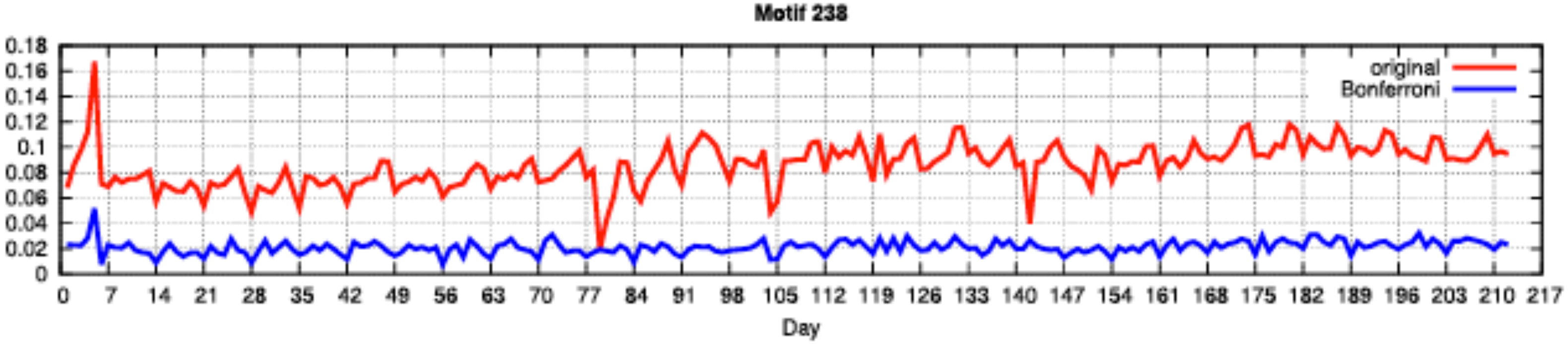
110




Motif 110

238




Motif 238

21 July, 2018

Interbank market: a study of the e-MID database and an agent based model

- electronic transactions between 254 Italian banks;

- transactions are transparent;

- time period from January 1999 to December 2009;

- information about the "aggressor" (lender or borrower);

- overnight and overnight-long credit relationships;

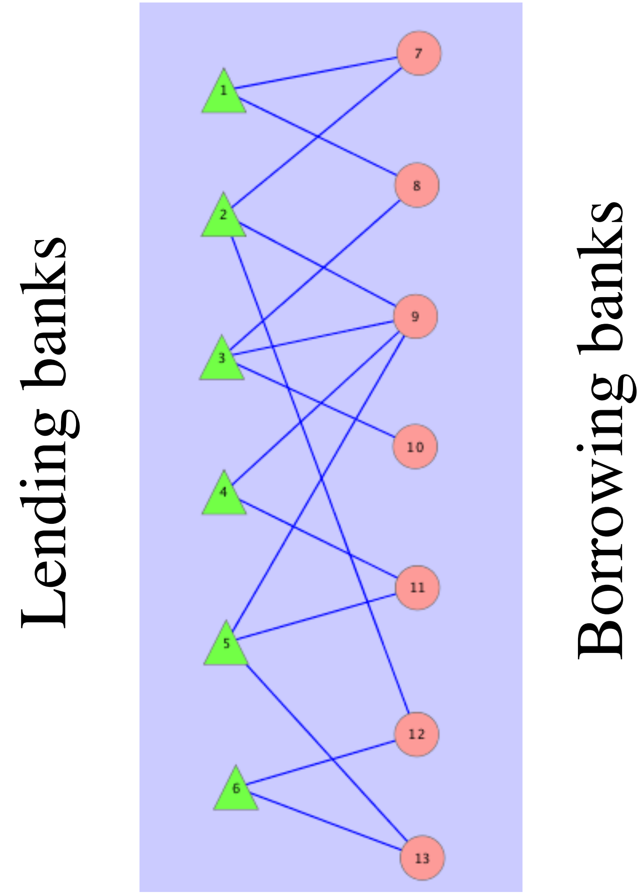
- data analyzed in 3-maintainance periods (1-maintainance period is about 23 trading days usually close to q1 calendar month).

¶Hatzopoulos, V., Iori, G., Mantegna, R. N., Miccichè, S., & Tumminello, M. (2015). Quantifying preferential trading in the e-MID interbank market. *Quantitative Finance*, *15*(4), 693-710.

¶Iori, G., Mantegna, R. N., Marotta, L., Miccichè, S., Porter, J., & Tumminello, M. (2015). Networked relationships in the e-MID Interbank market: A trading model with memory. Journal of Economic Dynamics and Control, 50, 98-116.

Empirical analyses[¶] of the e-MID database show evidence of the networked nature of the interbank market

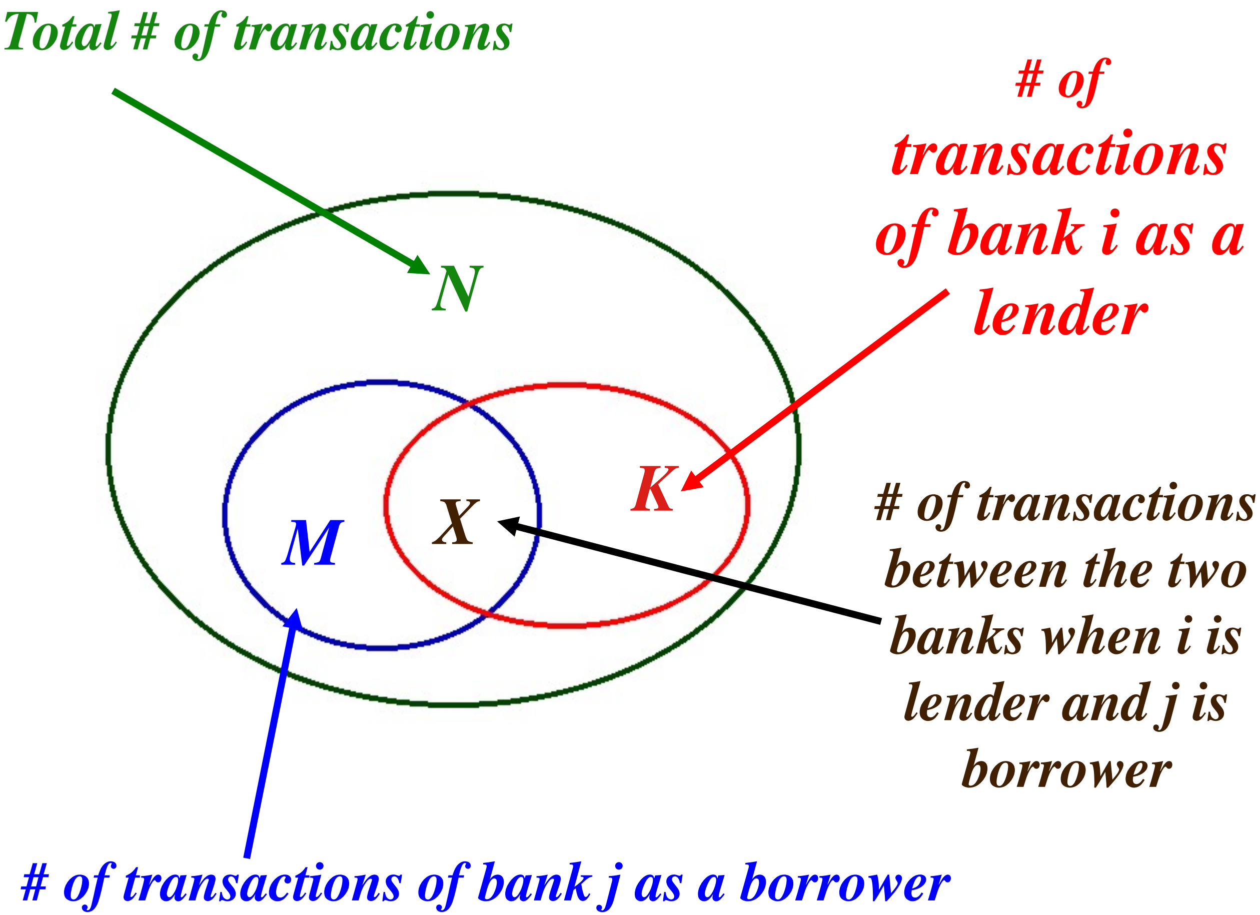
Lender aggressor or borrower aggressor transactions



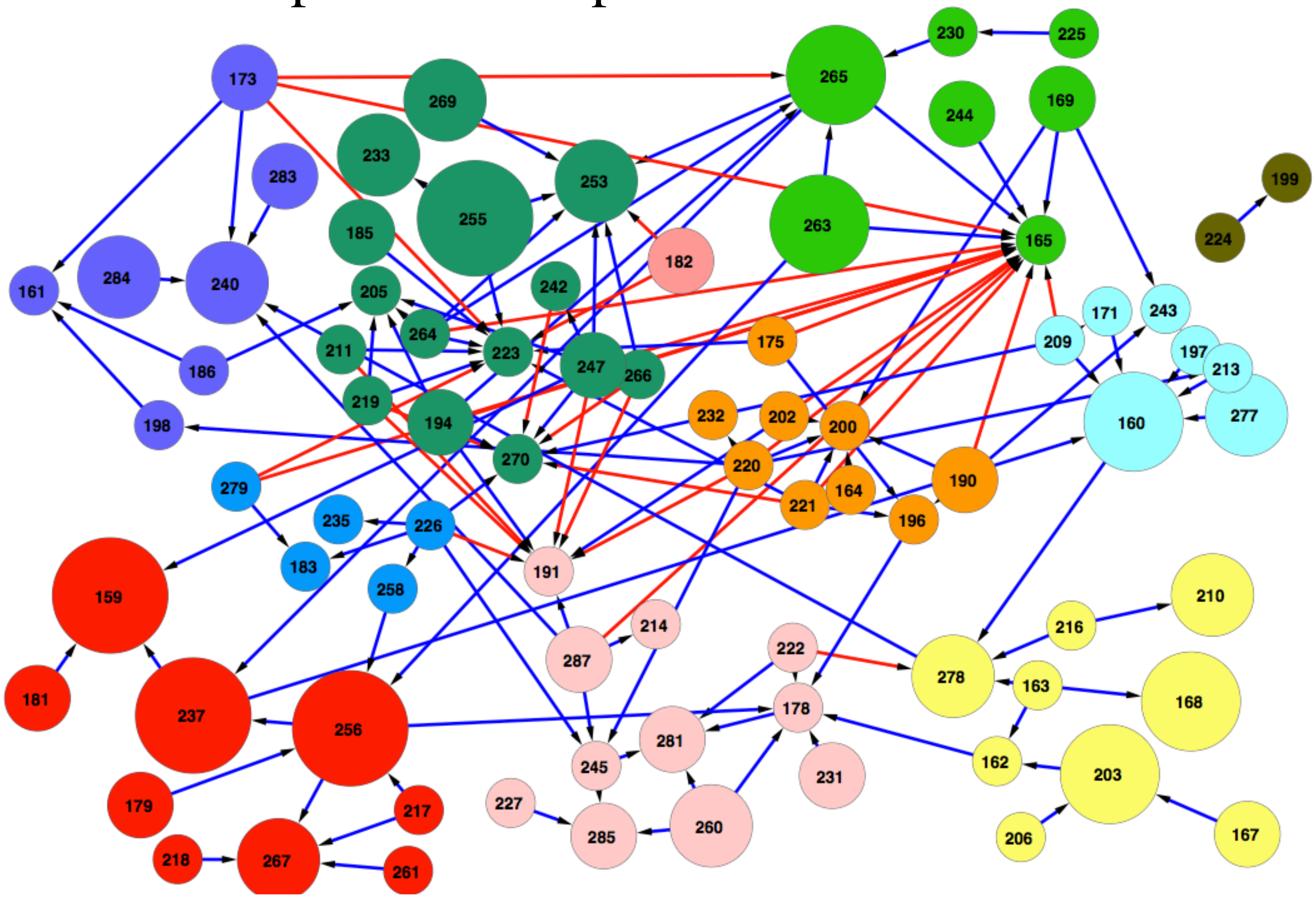
We performed a statistical validation of the over-expression and under-expression of repeated credit transactions.

[¶] Hatzopoulos, V., Iori, G., Mantegna, R. N., Miccichè, S., & Tumminello, M. (2015). Quantifying preferential trading in the e-MID interbank market. *Quantitative Finance*, *15*(4), 693-710.

We statistically validate the relationships between banks i (*lender*) and j (*borrower*) over a given time period.

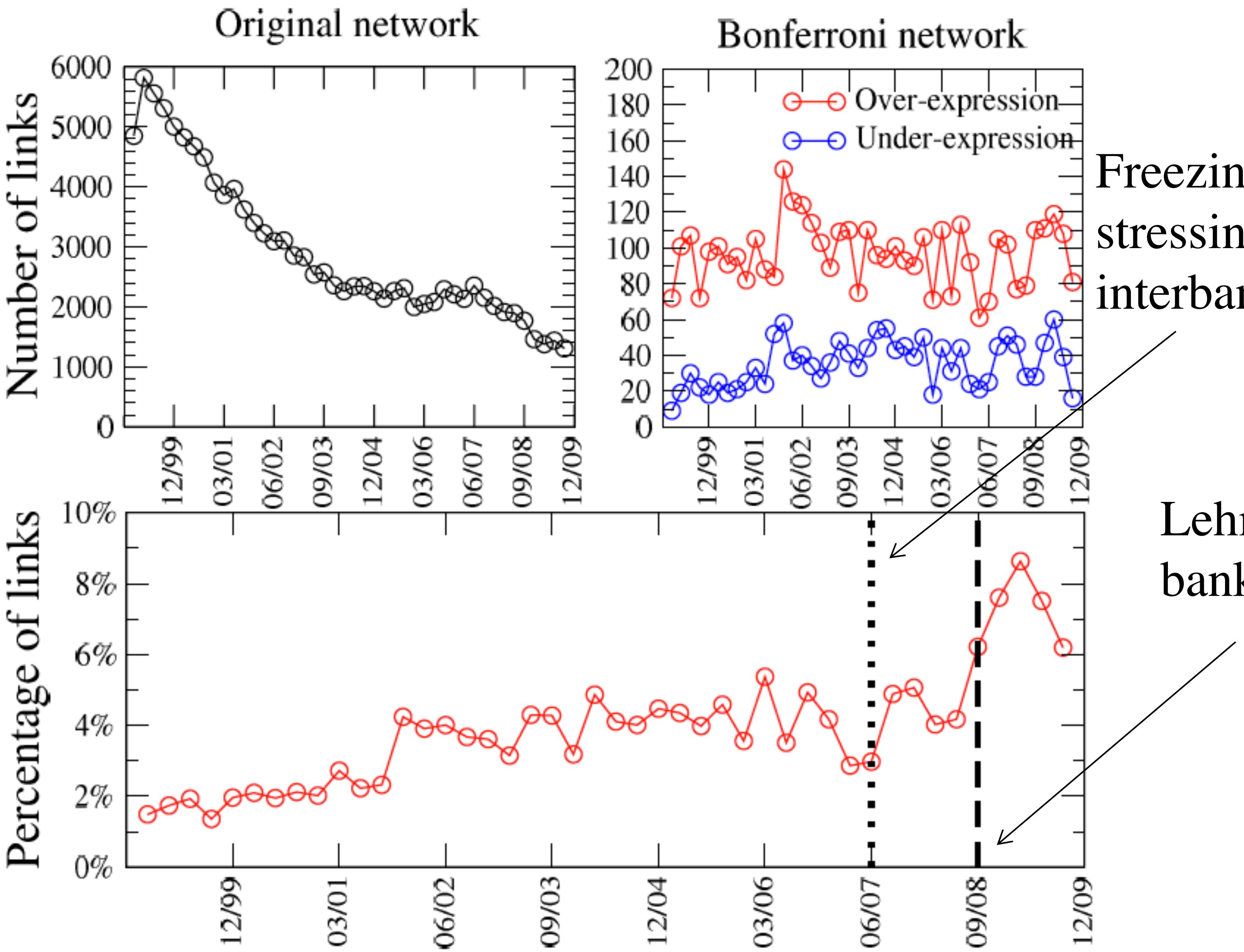

Total # of transactions
# of transactions of bank i as a lender
N
K
M
X
# of transactions between the two banks when i is lender and j is borrower
# of transactions of bank j as a borrower

Lender aggressor transactions. Bonferroni network of the 3-maintenance period 10-Sep-2008 / 09-Dec-2008



The different colors indicate the node membership to the partitions detected by using the Radatool algorithm (unweighted option). Red links are under-expressed links, while blue links are over-expressed ones.

Lender-aggressor dataset



Freezing (or stressing) of the interbank market

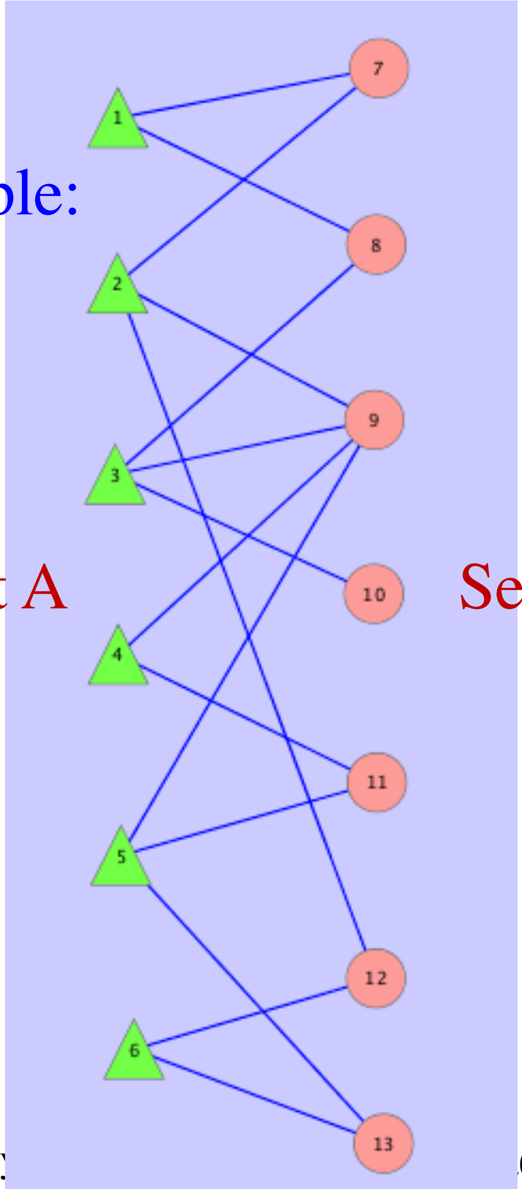
Lehman's bankruptcy.

In many cases investigated networks are projected networks of bipartite networks

Lending banks

Packages

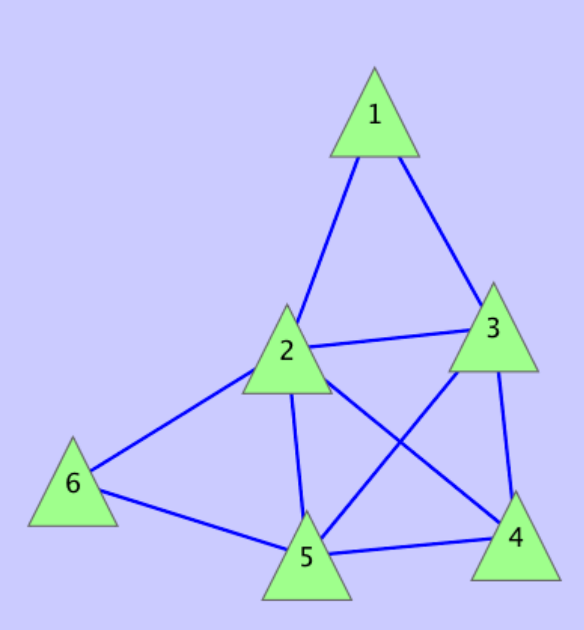
Example:



Set A

Set B

Projected network of lending banks



The investigated system concerns syndicated loans.

The database is the DealScan database of Thomson Reuters

In community detection one needs a "fitness" quantity to be minimized or maximized.

In network science a very popular measure is modularity.

By calling c_{i} the group to which vertex i belong and by defining δ(m,n) as the Kronecker delta, the modularity is

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

with $B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$ obeying to the relation

$$\sum_j B_{ij} = \sum_j A_{ij} - \frac{k_i}{2m}\sum_j k_j = k_i - \frac{k_i}{2m} 2m = 0$$

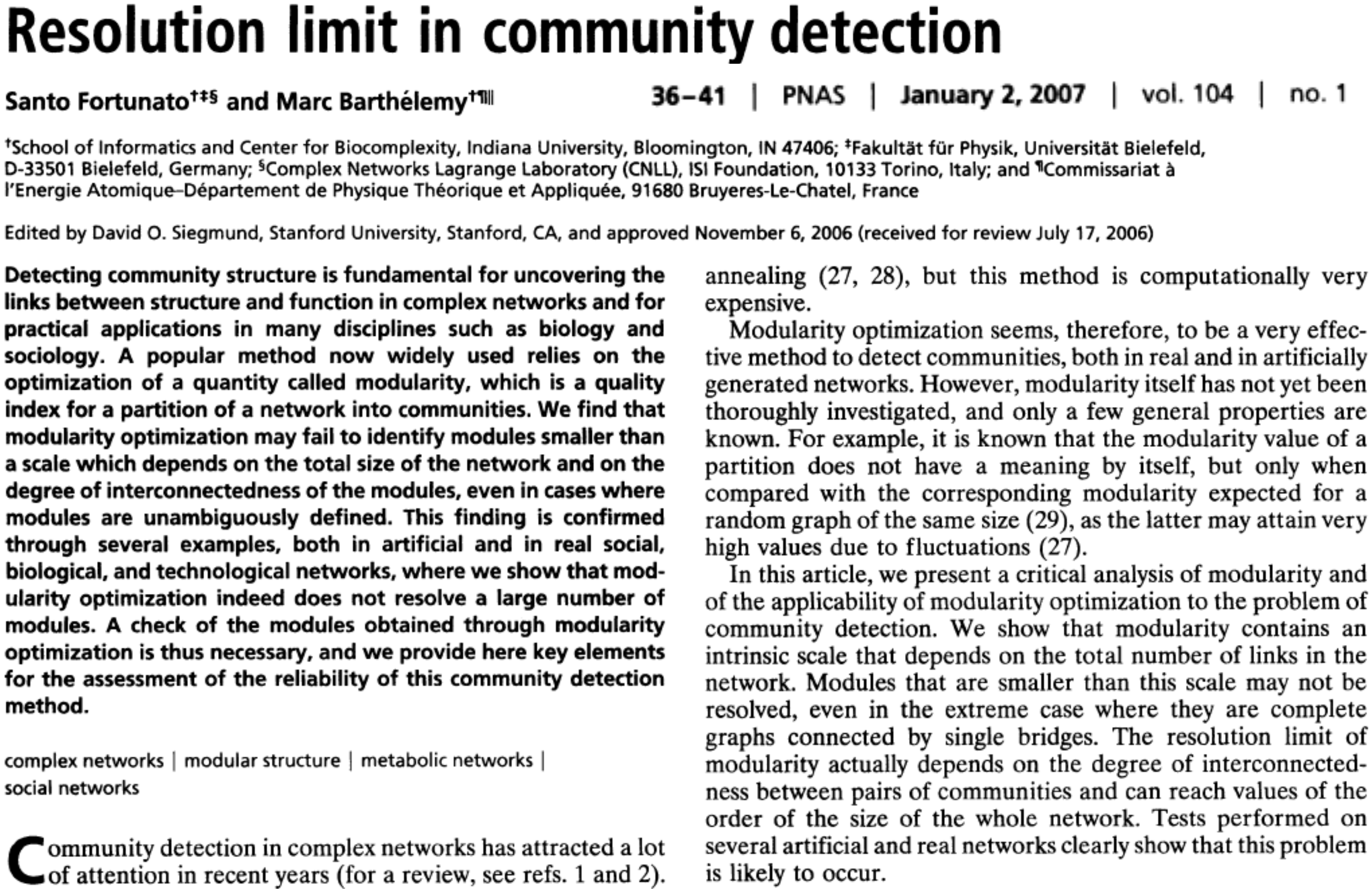
The modularity is comparing the degree of interconnection within a specific group with the degree of interconnection expected for a random connection of vertices preserving the vertex degree.

It should be noted that vertex i has degree k_i. There are $2m$ ends of edges in the entire network (m is the number of edges). Let us consider an edge starting from vertex i. The probability that the other end of the edge is vertex j is k_j/2m. Considering the degree of vertex i, the total expected number of edges between vertices i and j is

$$\frac{k_i k_j}{2m}$$

Newman, M.E. and Girvan, M., 2004. Finding and evaluating community structure in networks. Physical review E, 69(2), p.026113.

It has been observed that modularity has a resolution limit



Resolution limit in community detection

Santo Fortunato[†‡§] and Marc Barthélemy[†¶||]

36–41 | PNAS | January 2, 2007 | vol. 104 | no. 1

[†]School of Informatics and Center for Biocomplexity, Indiana University, Bloomington, IN 47406; [‡]Fakultät für Physik, Universität Bielefeld, D-33501 Bielefeld, Germany; [§]Complex Networks Lagrange Laboratory (CNLL), ISI Foundation, 10133 Torino, Italy; and [¶]Commissariat à l'Energie Atomique–Département de Physique Théorique et Appliquée, 91680 Bruyeres-Le-Chatel, France

Edited by David O. Siegmund, Stanford University, Stanford, CA, and approved November 6, 2006 (received for review July 17, 2006)

Detecting community structure is fundamental for uncovering the links between structure and function in complex networks and for practical applications in many disciplines such as biology and sociology. A popular method now widely used relies on the optimization of a quantity called modularity, which is a quality index for a partition of a network into communities. We find that modularity optimization may fail to identify modules smaller than a scale which depends on the total size of the network and on the degree of interconnectedness of the modules, even in cases where modules are unambiguously defined. This finding is confirmed through several examples, both in artificial and in real social, biological, and technological networks, where we show that modularity optimization indeed does not resolve a large number of modules. A check of the modules obtained through modularity optimization is thus necessary, and we provide here key elements for the assessment of the reliability of this community detection method.

complex networks | modular structure | metabolic networks | social networks

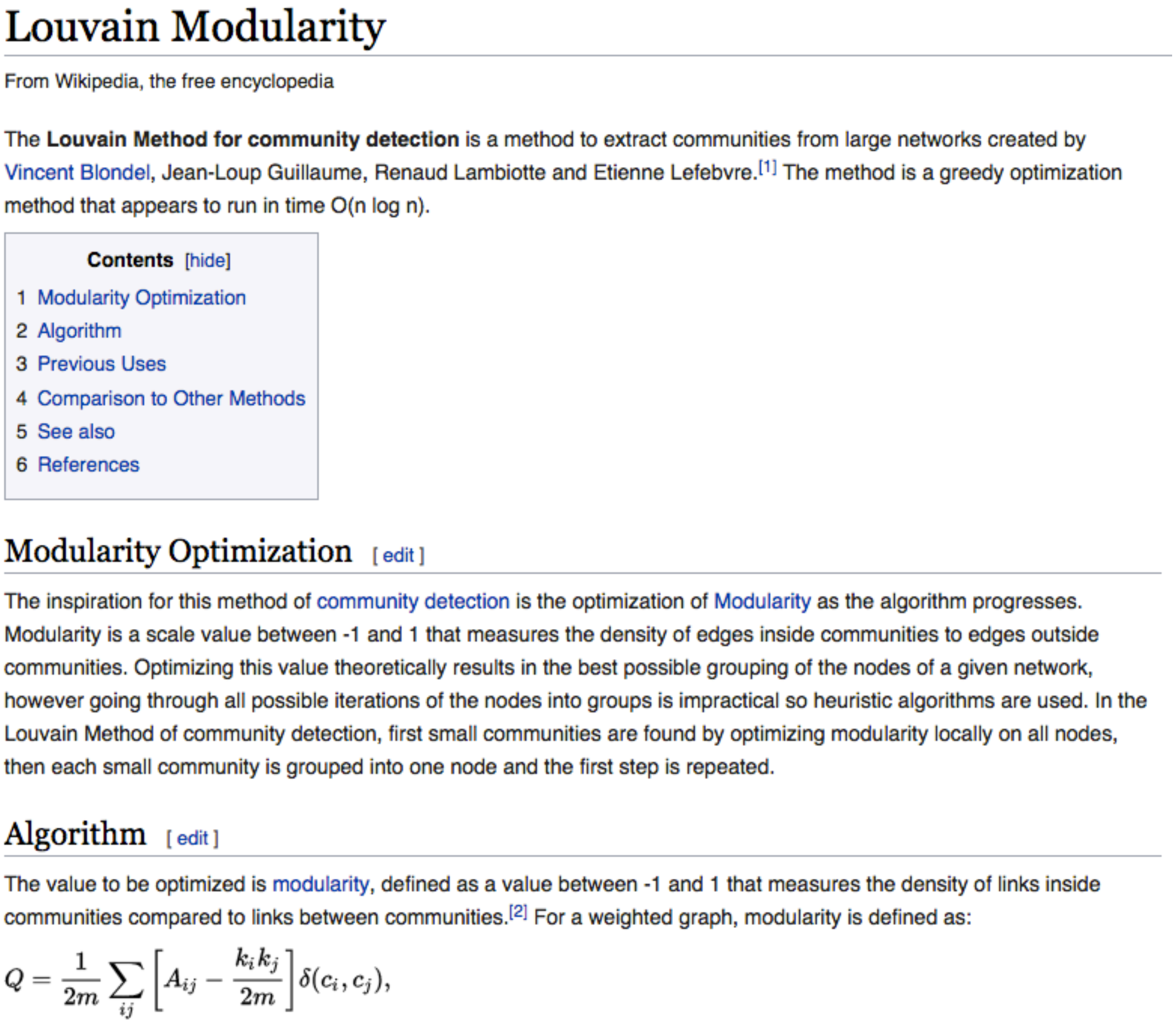
Community detection in complex networks has attracted a lot of attention in recent years (for a review, see refs. 1 and 2). annealing (27, 28), but this method is computationally very expensive.

Modularity optimization seems, therefore, to be a very effective method to detect communities, both in real and in artificially generated networks. However, modularity itself has not yet been thoroughly investigated, and only a few general properties are known. For example, it is known that the modularity value of a partition does not have a meaning by itself, but only when compared with the corresponding modularity expected for a random graph of the same size (29), as the latter may attain very high values due to fluctuations (27).

In this article, we present a critical analysis of modularity and of the applicability of modularity optimization to the problem of community detection. We show that modularity contains an intrinsic scale that depends on the total number of links in the network. Modules that are smaller than this scale may not be resolved, even in the extreme case where they are complete graphs connected by single bridges. The resolution limit of modularity actually depends on the degree of interconnectedness between pairs of communities and can reach values of the order of the size of the whole network. Tests performed on several artificial and real networks clearly show that this problem is likely to occur.

A different benchmark is therefore needed to have a better assessment of different community detection algorithms

A popular community detection algorithm based on modularity maximization: the Louvain algorithm



Louvain Modularity

From Wikipedia, the free encyclopedia

The **Louvain Method for community detection** is a method to extract communities from large networks created by Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre.[1] The method is a greedy optimization method that appears to run in time O(n log n).

Contents [hide]

Modularity Optimization [edit]

The inspiration for this method of community detection is the optimization of Modularity as the algorithm progresses. Modularity is a scale value between -1 and 1 that measures the density of edges inside communities to edges outside communities. Optimizing this value theoretically results in the best possible grouping of the nodes of a given network, however going through all possible iterations of the nodes into groups is impractical so heuristic algorithms are used. In the Louvain Method of community detection, first small communities are found by optimizing modularity locally on all nodes, then each small community is grouped into one node and the first step is repeated.

Algorithm [edit]

The value to be optimized is modularity, defined as a value between -1 and 1 that measures the density of links inside communities compared to links between communities.[2] For a weighted graph, modularity is defined as:

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

Blondel, V.D., Guillaume, J.L., Lambiotte, R. and Lefebvre, E., 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), p.P10008.

Rugged shape of the high-modularity region in the configuration space of network partitions.

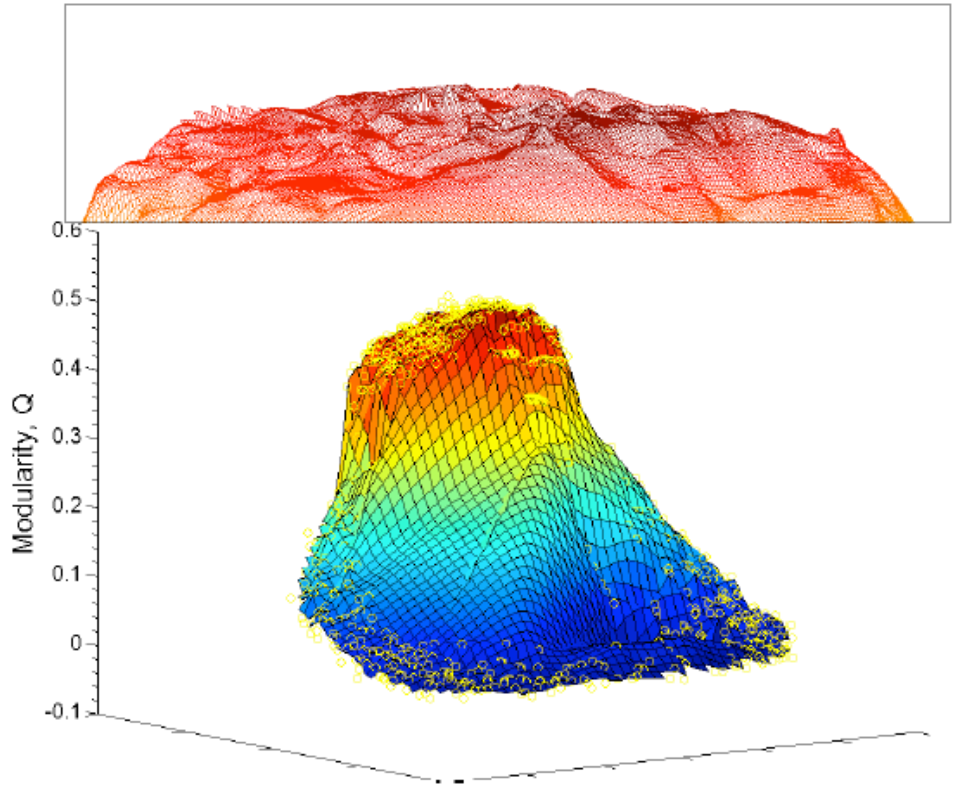


FIG. 3: (color online) The modularity function of a hierarchical random graph model [47], with $n = 256$ nodes arranged in a balanced hierarchy with assortative modules (see Appendix E), reconstructed from 1199 sampled partitions (circles), and its rugged high-modularity region (inset).

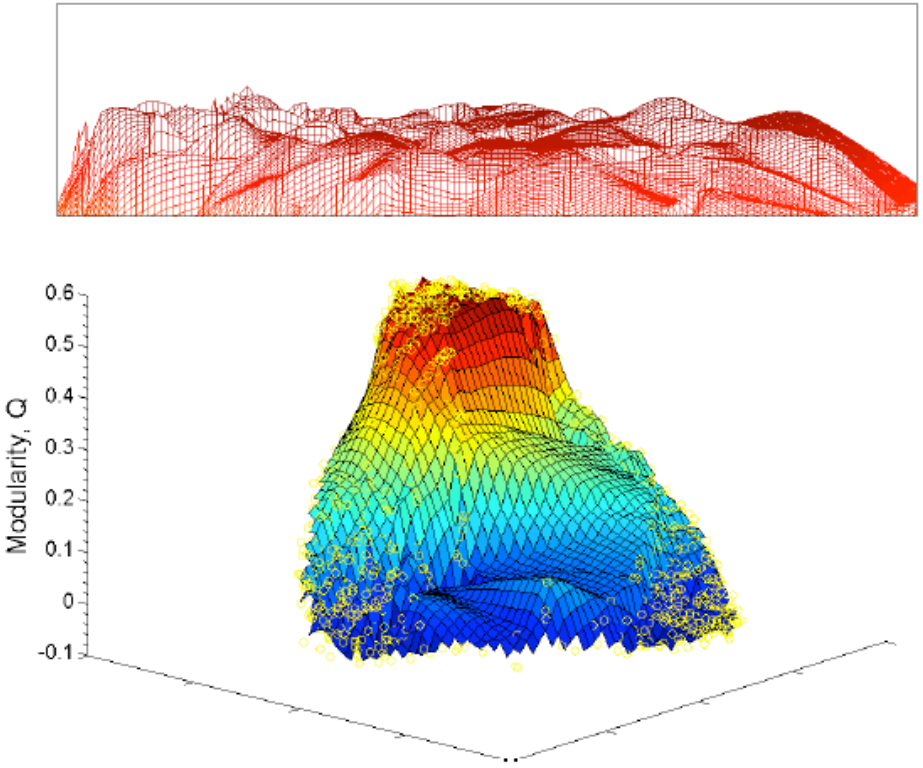


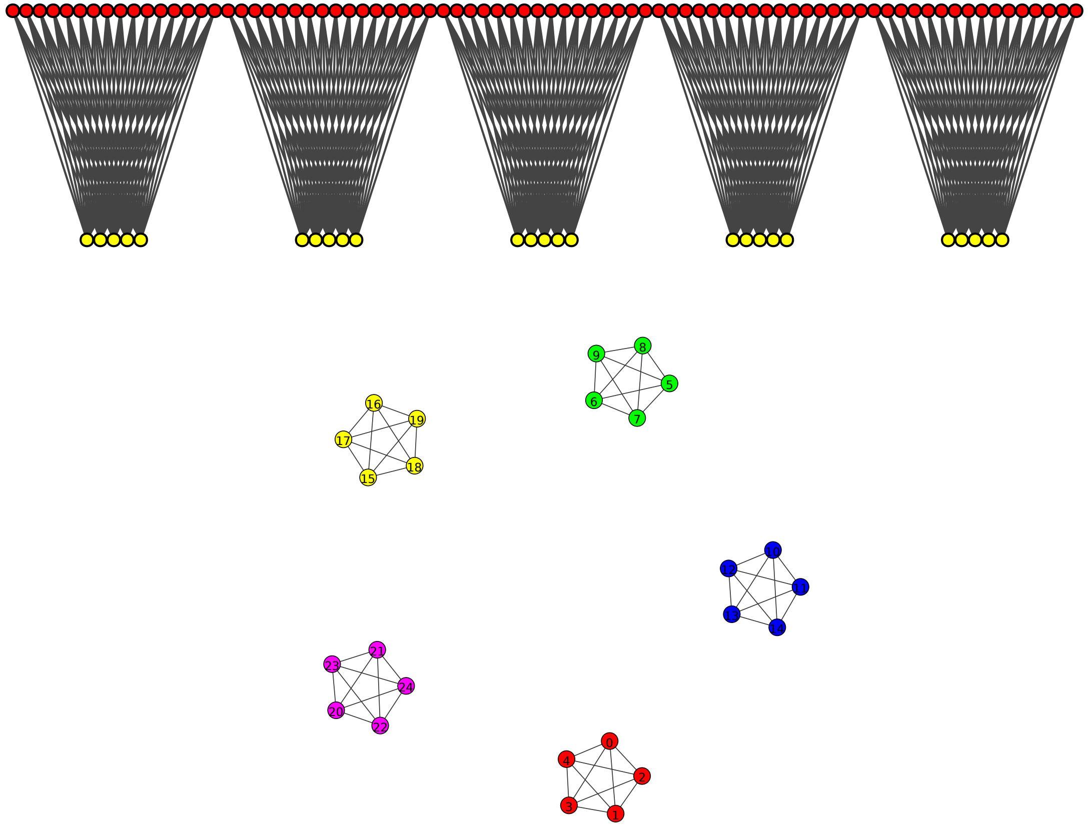
FIG. 4: (color online) The modularity function for the metabolic network of the spirochaete *Treponema pallidum* with $n = 482$ nodes (the largest component) and 1199 sampled partitions, showing qualitatively the same structure as we observed for hierarchical networks. The inset shows the rugged high-modularity region.

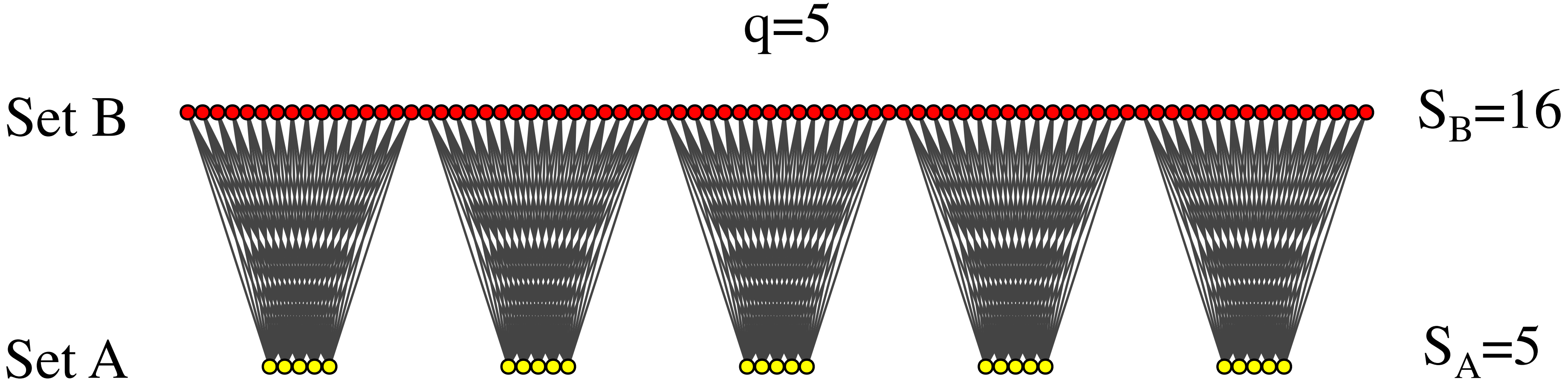
Good, Benjamin H., Yves-Alexandre de Montjoye, and Aaron Clauset. "Performance of modularity maximization in practical contexts." Physical Review E 81.4 (2010): 046106.

The use of network benchmarks in the comparative evaluation of different community detection algorithms

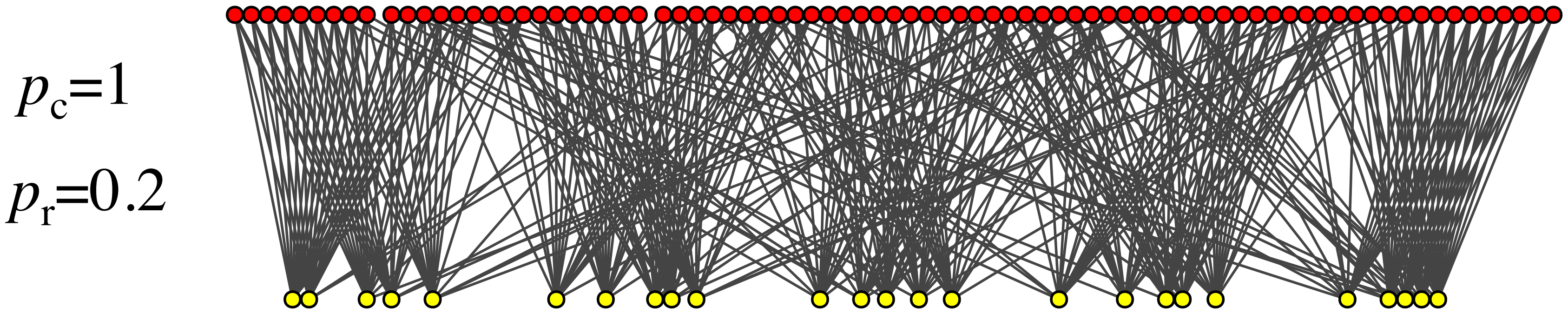
Lancichinetti, A., Fortunato, S. and Radicchi, F., 2008. Benchmark graphs for testing community detection algorithms. Physical review E, 78(4), p.046110.

We devised our own network benchmark designed to investigate community detection in projected network obtained starting from a bipartite network



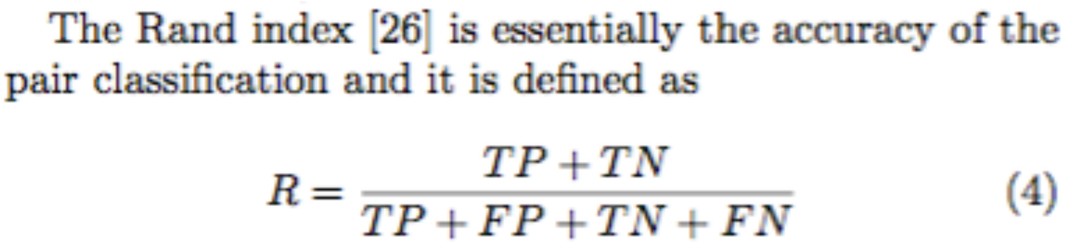


We obtain different realizations of the benchmark by varying q, S_A, S_B, the probability of coverage p_c, and the probability of re-assignment p_r



Comparing partitions

Partition 1 Partition 2

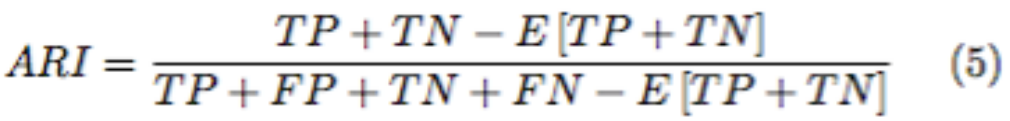


The Rand index [26] is essentially the accuracy of the pair classification and it is defined as

$$R = \frac{TP + TN}{TP + FP + TN + FN} \qquad (4)$$

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336), pp.846-850.

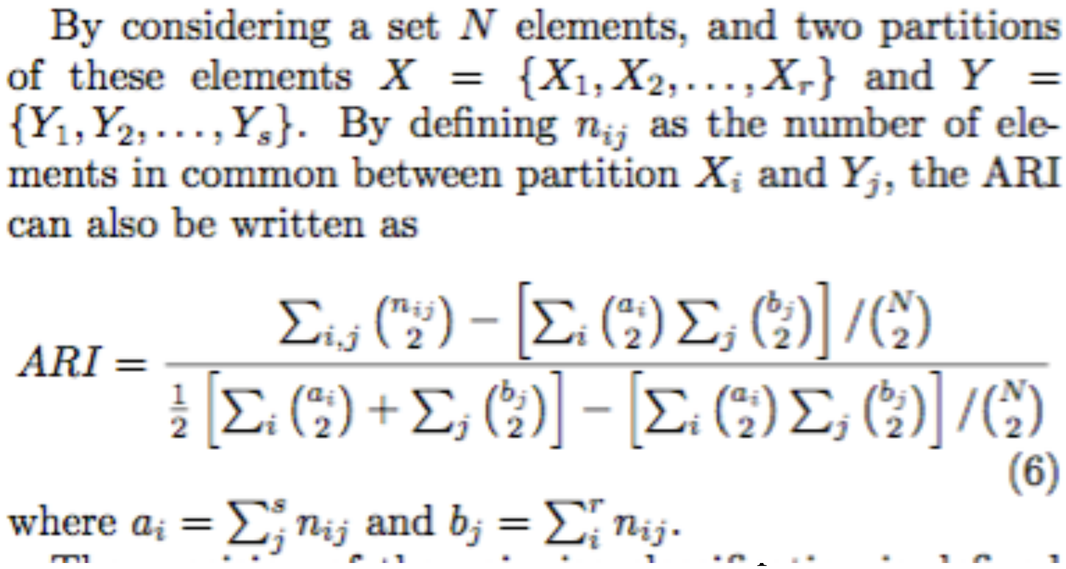
The adjusted Rand index is defined as



$$ARI = \frac{TP + TN - E\left[TP + TN\right]}{TP + FP + TN + FN - E\left[TP + TN\right]} \qquad (5)$$

Hubert, L. and Arabie, P., 1985. Comparing partitions. Journal of classification, 2(1), pp.193-218.

or equivalently

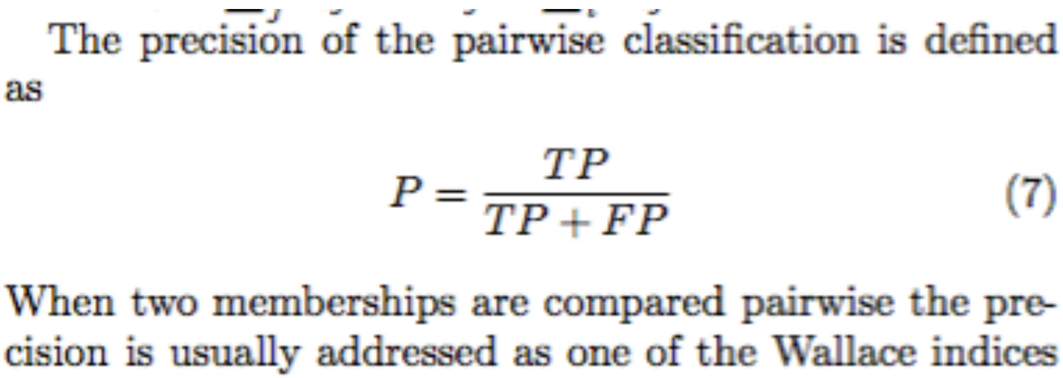


By considering a set N elements, and two partitions of these elements $X = \{X_1, X_2, \ldots, X_r\}$ and $Y = \{Y_1, Y_2, \ldots, Y_s\}$. By defining n_{ij} as the number of elements in common between partition X_i and Y_j, the ARI can also be written as

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{N}{2}} \qquad (6)$$

where $a_i = \sum_j^s n_{ij}$ and $b_j = \sum_i^r n_{ij}$.

Precision of detecting a pair of nodes in two distinct partitions

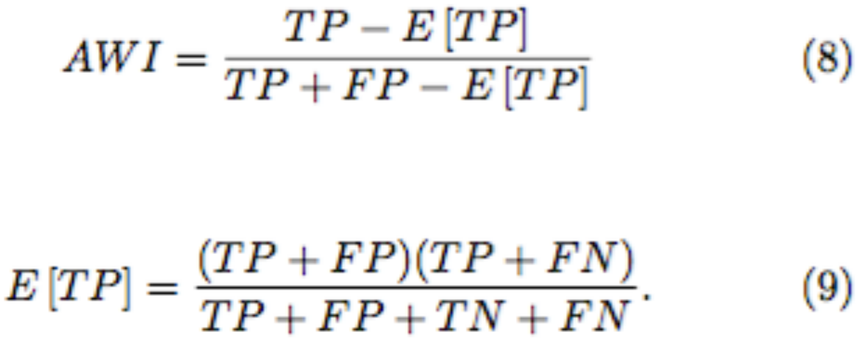


The precision of the pairwise classification is defined as

$$P = \frac{TP}{TP + FP} \qquad (7)$$

When two memberships are compared pairwise the precision is usually addressed as one of the Wallace indices

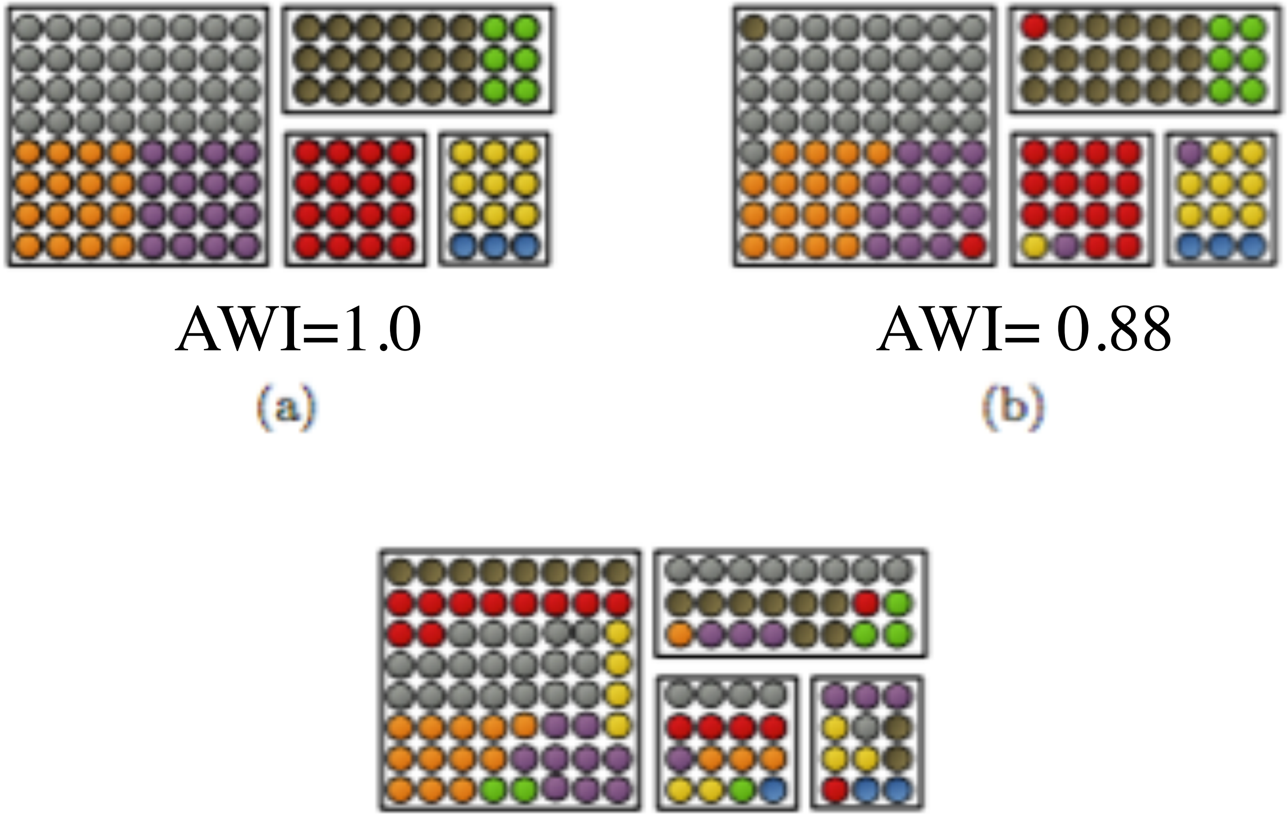
Wallace, D.L., 1983. Comment. Journal of the American Statistical Association, 78(383), pp.569-576.



$$AWI = \frac{TP - E[TP]}{TP + FP - E[TP]} \qquad (8)$$

where

$$E[TP] = \frac{(TP + FP)(TP + FN)}{TP + FP + TN + FN}. \qquad (9)$$



AWI=1.0

(a)

AWI= 0.88

(b)

AWI=0.03

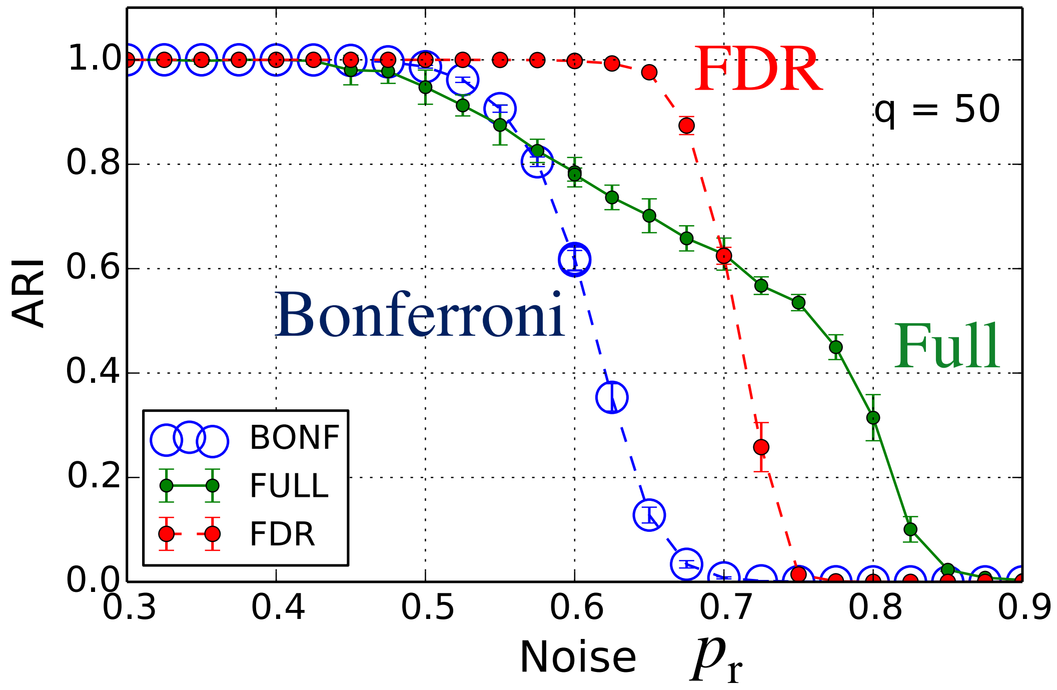
(c)

Partition 2 (colors)
cluster light gray: 32 el.
cluster gray: 18 el.
cluster orange: 16 el.
cluster purple: 16 el.
cluster red: 16 el.
cluster yellow: 9 el.
cluster green: 6 el.
cluster blue: 3 el.

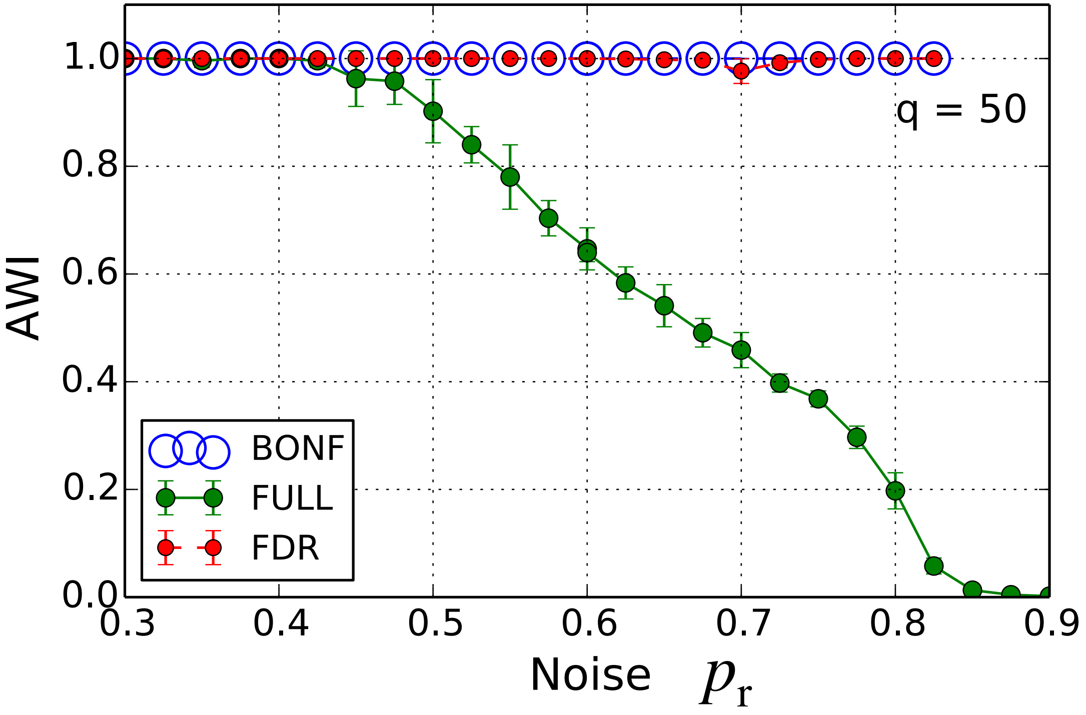
Partition 1 (boxes)
cluster I: 64 elements
cluster II: 24 elements
cluster III: 16 elements
cluster IV: 12 elements

FIG. 2: Three examples of comparison of a considered partition (membership of nodes indicated by different colors) with a reference partition (membership of nodes indicated by their position in different boxes). In the example a system of 116 nodes has 4 communities of different size in the reference partition (see four boxes with 64, 24, 16, and 12 nodes) and 8 communities of different size in the considered partition. This second partition is indicated by the colors of nodes. We have light gray (32 nodes), gray (18), orange (16), purple (16), red (16), yellow (9), green (6), and blue (3) groups. In the three examples the AWI assumes the values: (a) $AWI = 1.0$, (b) $AWI = 0.88$, and (c) $AWI = 0.03$

Numerical simulations of a benchmark with q=50, S_A=50, and S_B=50, p_c=0.8



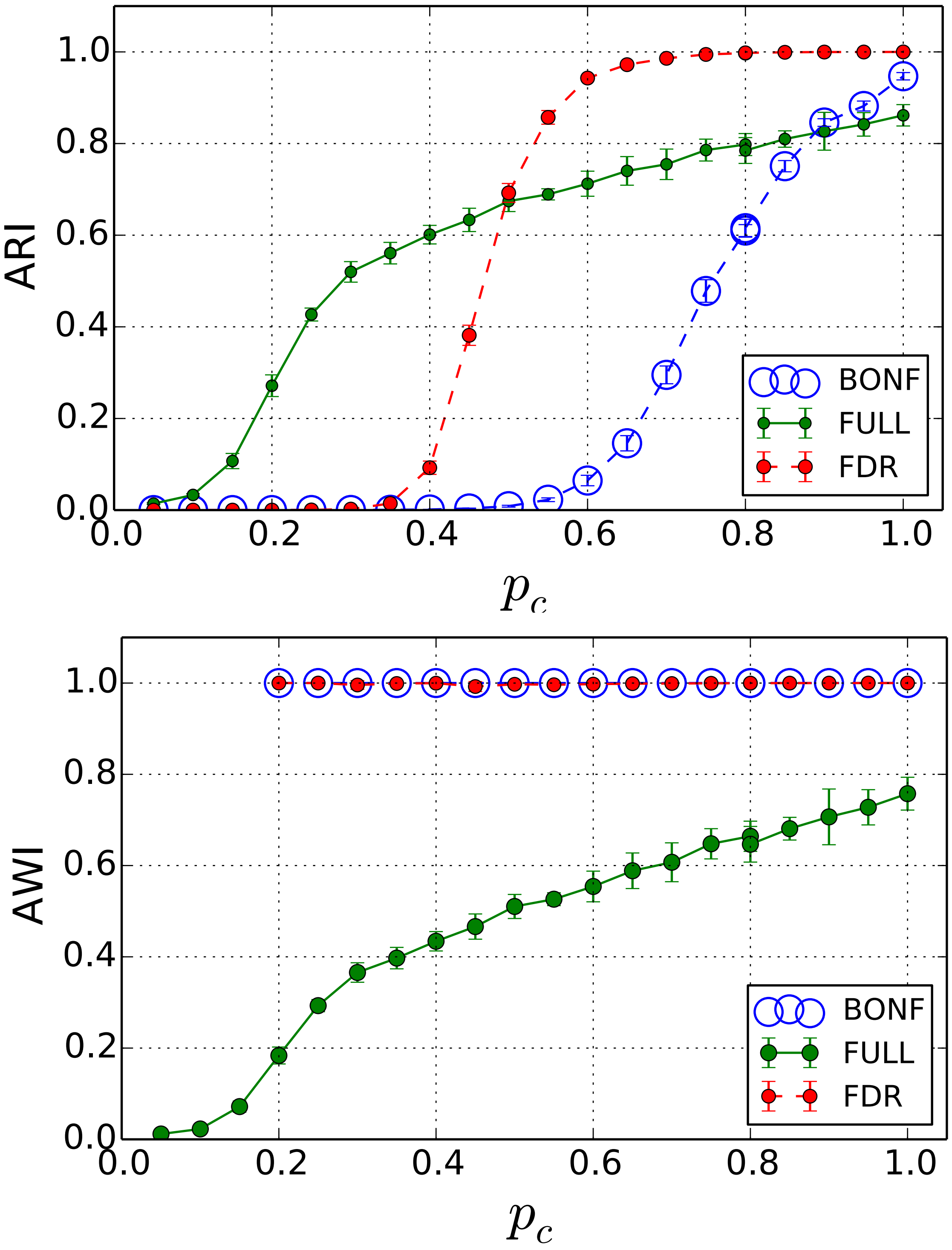
Investigation of the role of the probability of re-assignment p_r mimicking the presence of noise or errors



ARI and AWI are computed with respect to the reference partition

Numerical simulations of a benchmark with q=50, S_A=50, and S_B=50, p_r=0.6

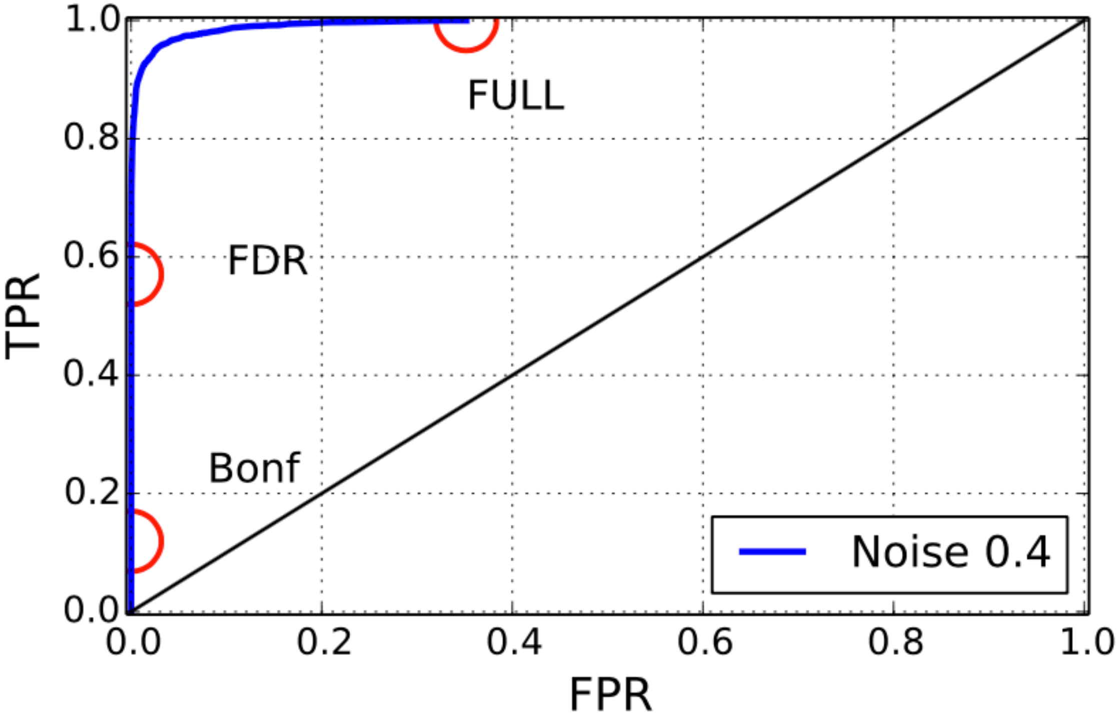
Investigation of the role of the probability of coverage p_c mimicking the completenes of the coverage of the available records



Receiver Operating Charactristic (ROC) for pairs classification

q=20 S_A=20 S_B=20 p_c=0.8 p_r=0.4

$$TRP = \frac{TP}{TP + FN}$$



The methodology has

$$FPR = \frac{FP}{FP + TN}$$

high $\text{Precision} = \dfrac{TP}{TP + FP}$

but might have low $\text{Accuracy} = \dfrac{TP + TN}{FP + TP + FN + TN}$

Robustness of partitions obtained with the Louvain algorithm in a real system

First real system: co-authorship network lcc comprising 13861 authors 19466 papers

We perform 1000 times the community detection algorithm finding partitions with modularity ranging from 0.864 to 0.867 and we select among the 1000 partitions the ten partitions with top values of modularity Tp_i *(modularity ranging from 0.8666 to 0.8670)*

For the full network: $\mathrm{ARI}(\mathrm{Tp}_i,\mathrm{Tp}_j)$ is ranging from 0.59 to 0.71

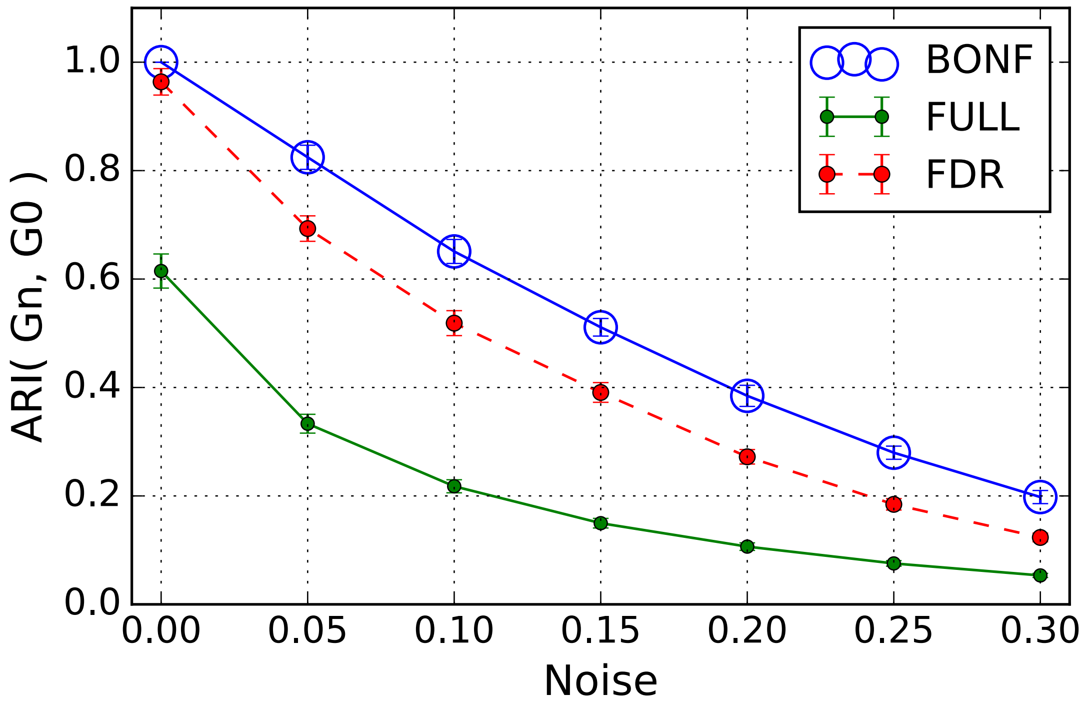
For the FDR network: $\mathrm{ARI}(\mathrm{Tp}_i,\mathrm{Tp}_j)$ is 1.00

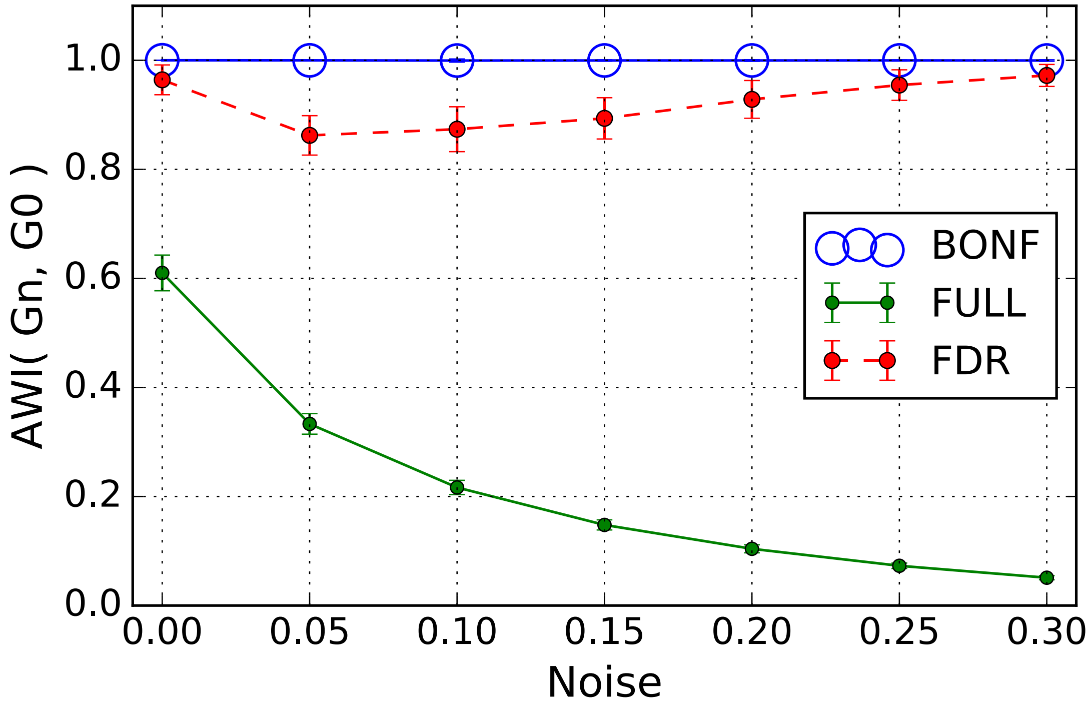
It is worth noting that the AWI is ranging from 0.57 to 0.66 indicating that the partitions obtained with FDR are not fully included in the partitions of the full network

Robustness to errors or misclassification in real networks p_r

Co-authorship network (lcc) comprising
13861 authors
19466 papers

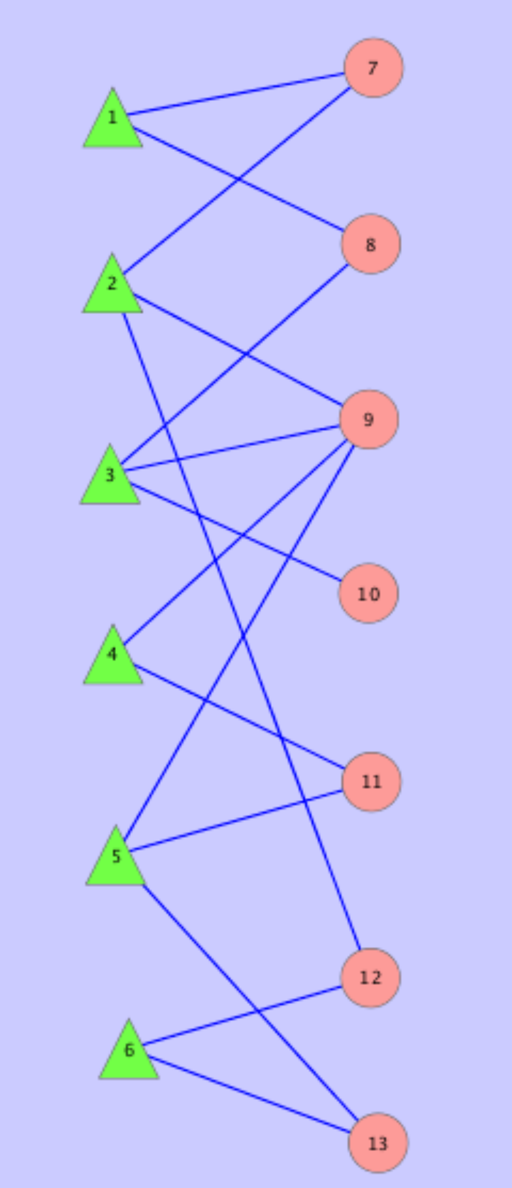
G_0 is the partition of highest modularity and for each value of pr we perform 100 different realizations.





Detecting categorical decisions of a heterogeneous system

Investors



buy decision at day i

sell decision at day i

buy-sell decision at day i

buy decision at day i+1

sell decision at day i+1

buy-sell decision at day i+1

....

Tumminello, M., Lillo, F., Piilo, J. and Mantegna, R.N., 2012. Identification of clusters of investors from their real trading activity in a financial market. New Journal of Physics, 14(1), p.013041.

Multi-link statistically validated network

Each investor can participate to 9 different types of co-occurrences of the 3 states Buy, Sell and BuySell with any other investor

Two investors participating to the Bonferroni or in the FDR network will be characterized by a link describing one, or more than one, of the 9 possible validated co-occurrences.

Multi-link statistically validated network

		Investor 2		
		Buy	Sell	BuySell
Investor 1	Buy	1	0	0
	Sell	0	1	0
	BuySell	0	0	0

With this matrix representation we mean that the FDR link indicates co-occurrence of the **Buy I_1** – **Buy I_2** and **Sell I_1** – **Sell I_2** activities

The number of different co-occurrence combinations is 2^9=512

During 1998-2003 14,735 Nokia investors did more than 20 transactions

The **Bonferroni** network comprises 3,118 investors which are connected by 36,664 multi-links

The **FDR** network comprises 10,435 investors which are connected by 330,404 multi-links

More than 99% of multi-links belong to just 9 different co-occurrence combinations

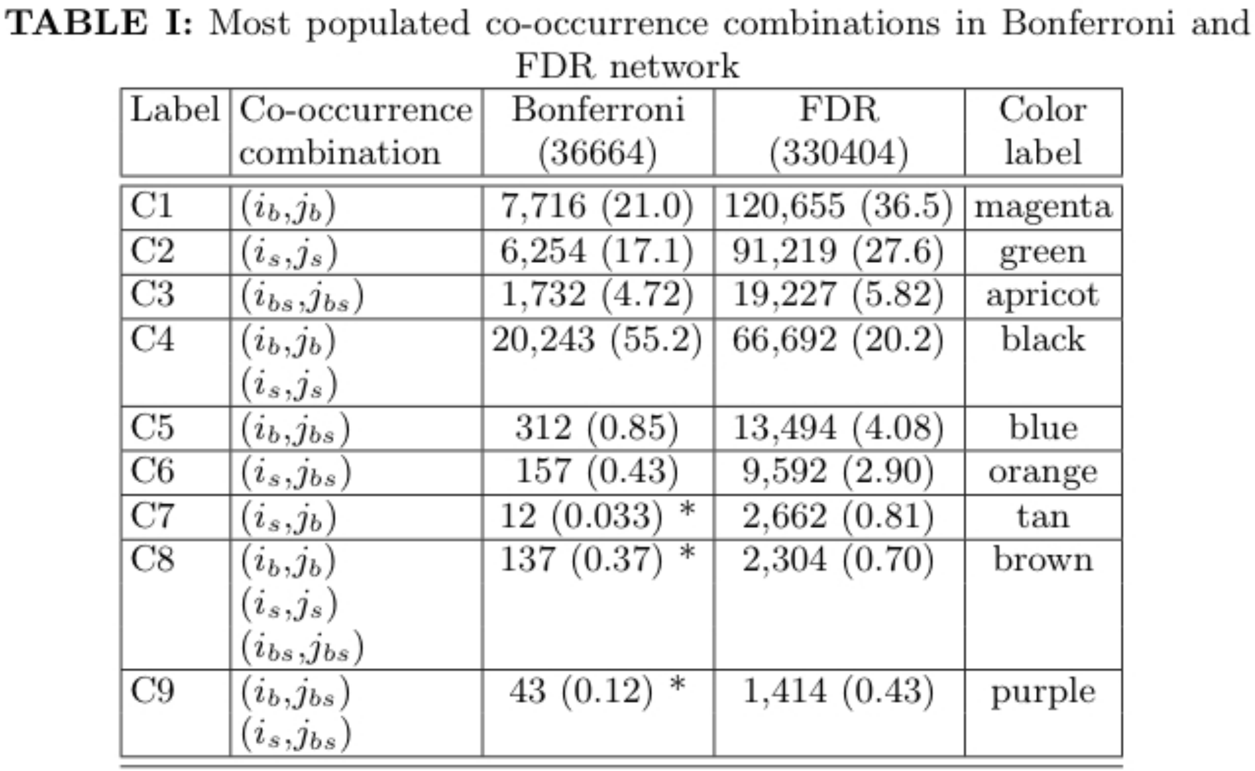


TABLE I: Most populated co-occurrence combinations in Bonferroni and FDR network

Label	Co-occurrence combination	Bonferroni (36664)	FDR (330404)	Color label
C1	(i_b, j_b)	7,716 (21.0)	120,655 (36.5)	magenta
C2	(i_s, j_s)	6,254 (17.1)	91,219 (27.6)	green
C3	(i_{bs}, j_{bs})	1,732 (4.72)	19,227 (5.82)	apricot
C4	(i_b, j_b) (i_s, j_s)	20,243 (55.2)	66,692 (20.2)	black
C5	(i_b, j_{bs})	312 (0.85)	13,494 (4.08)	blue
C6	(i_s, j_{bs})	157 (0.43)	9,592 (2.90)	orange
C7	(i_s, j_b)	12 (0.033) *	2,662 (0.81)	tan
C8	(i_b, j_b) (i_s, j_s) (i_{bs}, j_{bs})	137 (0.37) *	2,304 (0.70)	brown
C9	(i_b, j_{bs}) (i_s, j_{bs})	43 (0.12) *	1,414 (0.43)	purple

Top multi-links of the FDR network of investors

When $\theta=0.01$ we detect a network of **10435 investors** connected by **330404 multi-links**. The most common kinds of multi-links are

$$\begin{array}{c|ccc} & \mathbf{B} & \mathbf{S} & \mathbf{BS} \\ \hline B & \mathbf{1} & 0 & 0 \\ S & 0 & 0 & 0 \\ BS & 0 & 0 & 0 \end{array}$$

1 120655
magenta

$$\begin{array}{c|ccc} & \mathbf{B} & \mathbf{S} & \mathbf{BS} \\ \hline B & 0 & 0 & 0 \\ S & 0 & \mathbf{1} & 0 \\ BS & 0 & 0 & 0 \end{array}$$

2 91219
light green

$$\begin{array}{c|ccc} & \mathbf{B} & \mathbf{S} & \mathbf{BS} \\ \hline B & \mathbf{1} & 0 & 0 \\ S & 0 & \mathbf{1} & 0 \\ BS & 0 & 0 & 0 \end{array}$$

3 66692
black

$$\begin{array}{c|ccc} & \mathbf{B} & \mathbf{S} & \mathbf{BS} \\ \hline B & 0 & 0 & 0 \\ S & 0 & 0 & 0 \\ BS & 0 & 0 & \mathbf{1} \end{array}$$

4 19227
apricot

$$\begin{array}{c|ccc} & \mathbf{B} & \mathbf{S} & \mathbf{BS} \\ \hline B & 0 & 0 & \mathbf{1} \\ S & 0 & 0 & 0 \\ BS & 0 & 0 & 0 \end{array}$$

5 13494
blue

$$\begin{array}{c|ccc} & \mathbf{B} & \mathbf{S} & \mathbf{BS} \\ \hline B & 0 & 0 & 0 \\ S & 0 & 0 & \mathbf{1} \\ BS & 0 & 0 & 0 \end{array}$$

6 9592
orange

$$\begin{array}{c|ccc} & \mathbf{B} & \mathbf{S} & \mathbf{BS} \\ \hline B & 0 & 0 & 0 \\ S & \mathbf{1} & 0 & 0 \\ BS & 0 & 0 & 0 \end{array}$$

7 2662
tan

$$\begin{array}{c|ccc} & \mathbf{B} & \mathbf{S} & \mathbf{BS} \\ \hline B & \mathbf{1} & 0 & 0 \\ S & 0 & \mathbf{1} & 0 \\ BS & 0 & 0 & \mathbf{1} \end{array}$$

8 2304
brown

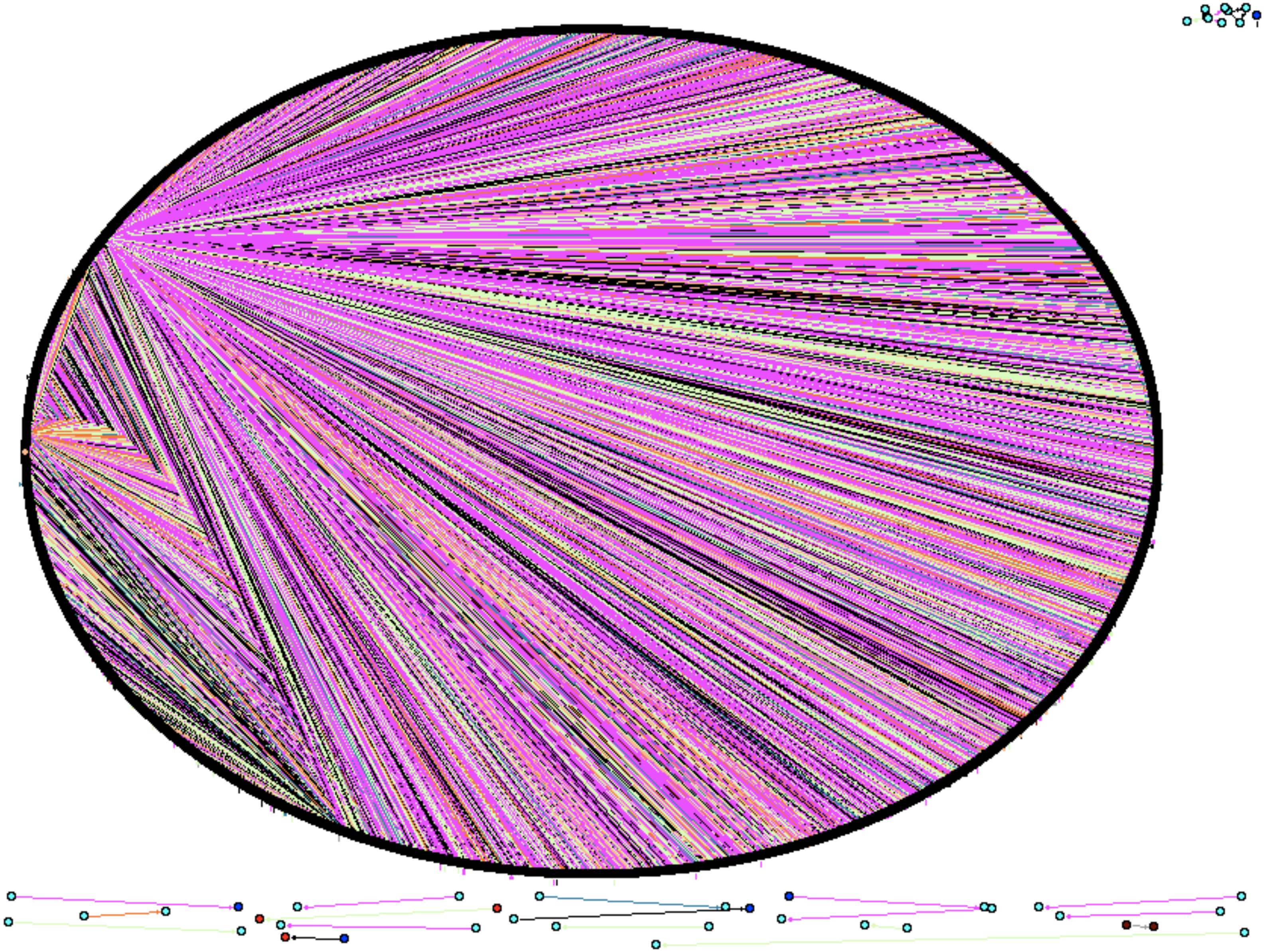
$$\begin{array}{c|ccc} & \mathbf{B} & \mathbf{S} & \mathbf{BS} \\ \hline B & 0 & 0 & 0 \\ S & 0 & 0 & 0 \\ BS & \mathbf{1} & \mathbf{1} & 0 \end{array}$$

9 1414
light purple

$$\begin{array}{c|ccc} & \mathbf{B} & \mathbf{S} & \mathbf{BS} \\ \hline B & \mathbf{1} & 0 & 0 \\ S & 0 & 0 & 0 \\ BS & 0 & \mathbf{1} & 0 \end{array}$$

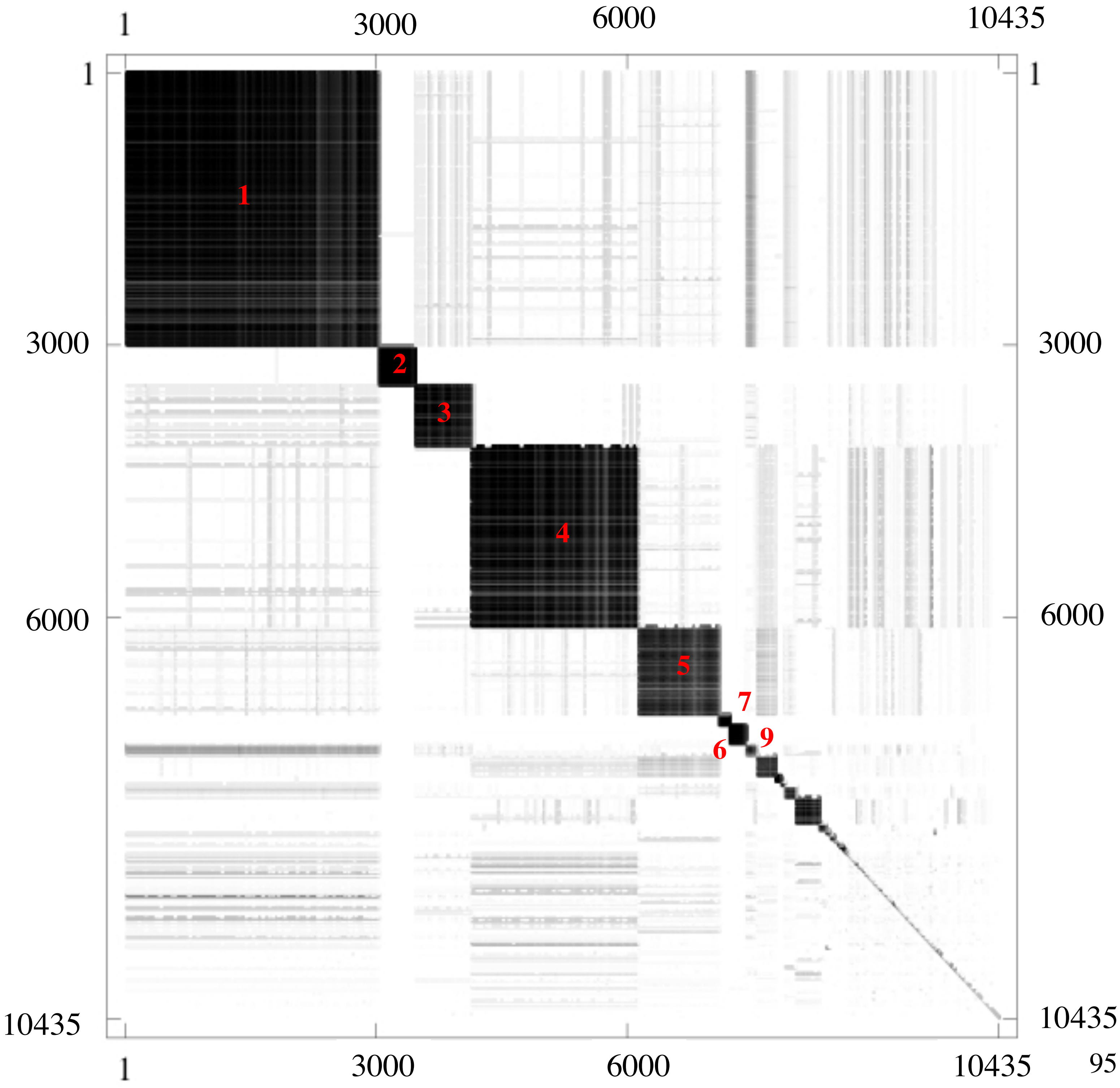
10 460
gray

The FDR network has a giant component covering almost entirely the set of investors (10392/10435)



We perform unsupervised community (cluster) detection in the statistically validated networks

Example: The Infomap partition of the FDR network of 10435 investors



Volume of the FDR network |Vb|+|Vs| =184 202 420 474

Cluster	Investors	% of Investors	Volume	% of Volume
1	3018	28.9	110 943 057 984	60.23
2	438	4.20	78 250 636	0.042
3	670	6.42	8 468 465 639	4.60
4	1995	19.1	43 432 935 550	23.58
5	963	9.23	9 198 936 014	4.99
6	126	1.21	3 187 754	0.0017
7	203	1.94	98 451 155	0.0534
8	127	1.22	70 571 657	0.0383
9	222	2.13	127 472 213	0.0692
10	70	0.671	1 424 940	0.0007
Clusters 1-10	7832	75.0	172 422 753 542	93.6
FDR	10435	100	184 202 420 474	100

FDR network

10435 Investors

Class	#
Gov. Investors	75
Companies	1472
Foreign inv.s	87
Households	8521
No profit	95
Fin. Institutions	185

Over-expression validation of vertex or link attributes

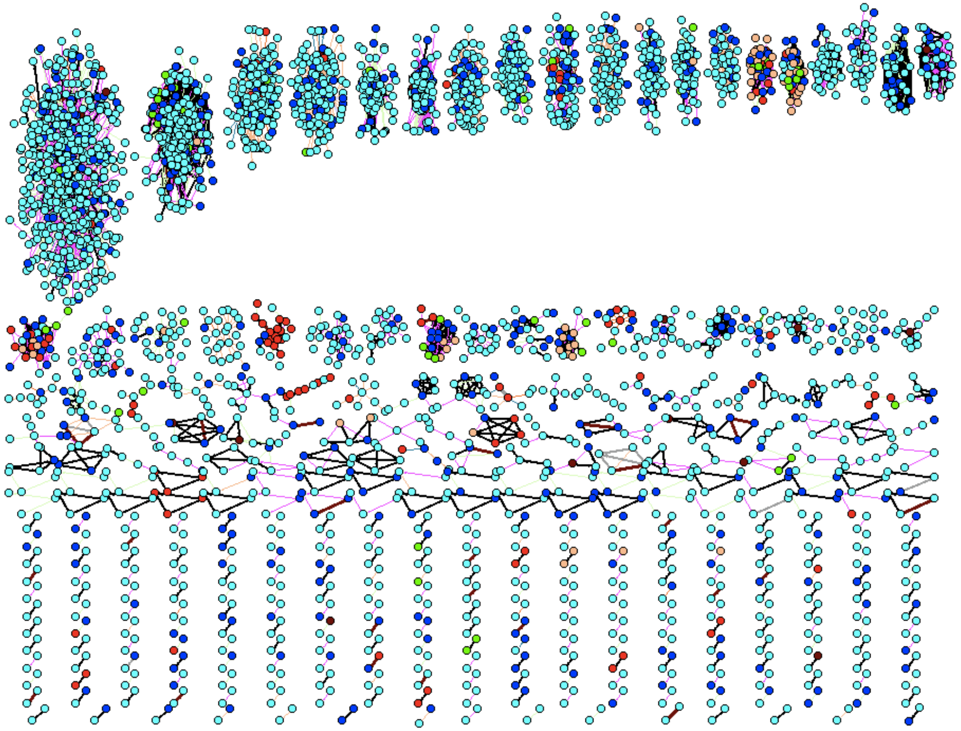
For a given set of elements (investors, links, etc) we count how many of them are present in our reference set. We count the same information also inside each subset of interest. For the sake of simplicity, let us focus on investor classes but similar conclusion applies for different attributes. For each subset **a** and for each investor class **k** we have the number $\mathbf{N_{a,k}}$ of investors of class **k** present in the subset **a**, the number $\mathbf{N_a}$ is the number of investors of subset **a,** $\mathbf{N_k}$ is the number of investors of class **k** in the subset and the number $\mathbf{N_n}$ is the number of investors in the reference set. The probability that **X** investors of subset **a** belongs to class **k** under a random null hypothesis is again given by the hypergeometric distribution $\mathbf{H(X|N_n,N_a,N_k)}$ and a *p*-value can therefore be associated to the observation of $\mathbf{N_{a,k}}$ occurrence.

Over-expression validation of vertex or link attributes

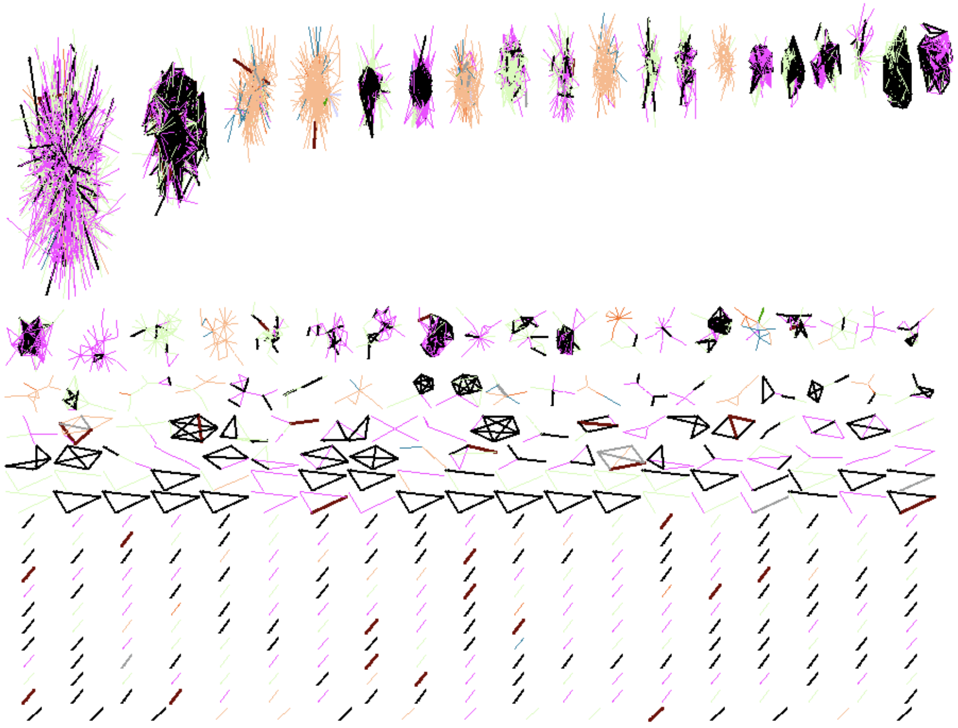
Again this is a multiple hypothesis test procedure and a Bonferroni threshold is set as **$0.01/N_{att}$** for each test of each partition, where **N_{att}** is the number of different attributes that are tested. In the example the number of different investor classes.

Michele Tumminello , Salvatore Miccichè , Fabrizio Lillo , Jan Varho , Jyrki Piilo and Rosario N Mantegna, Community characterization of heterogeneous complex systems J. Stat. Mech. (2011) P01019 doi: 10.1088/1742-5468/2011/01/P01019

Clusters of the Bonferroni network



focus on vertices



focus on links

● Companies

● Gov. inst.

● Financial inst.

● Non profit org.

● Foreign inst.

● Households

▬ B.B

▬ S.S

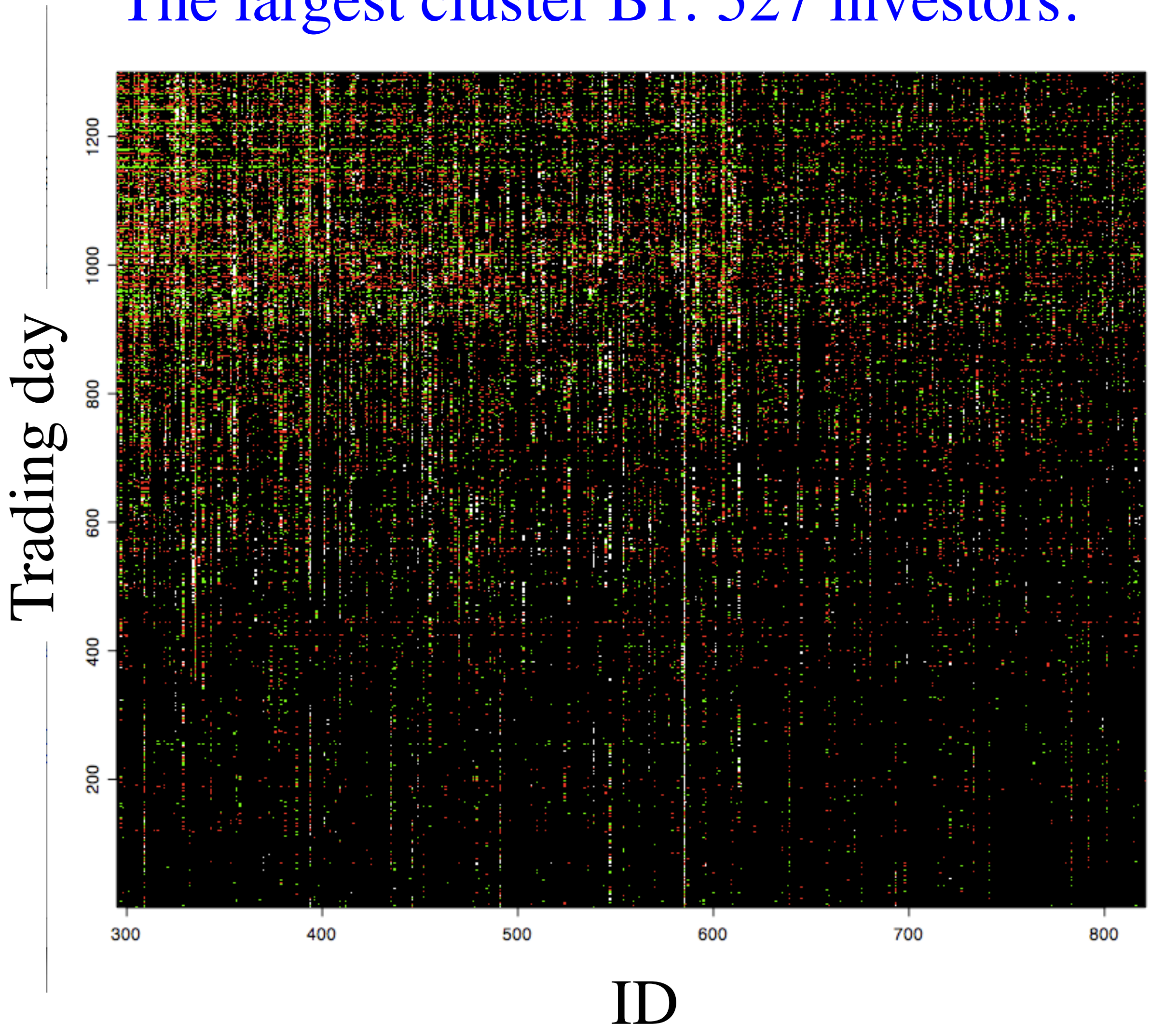
▬ B.B S.S

▬ BS.BS

▬ B.BS

Different clusters have different trading profiles and some of them have an over-expression of specific attributes of vertices and links

The largest cluster B1: 527 investors.

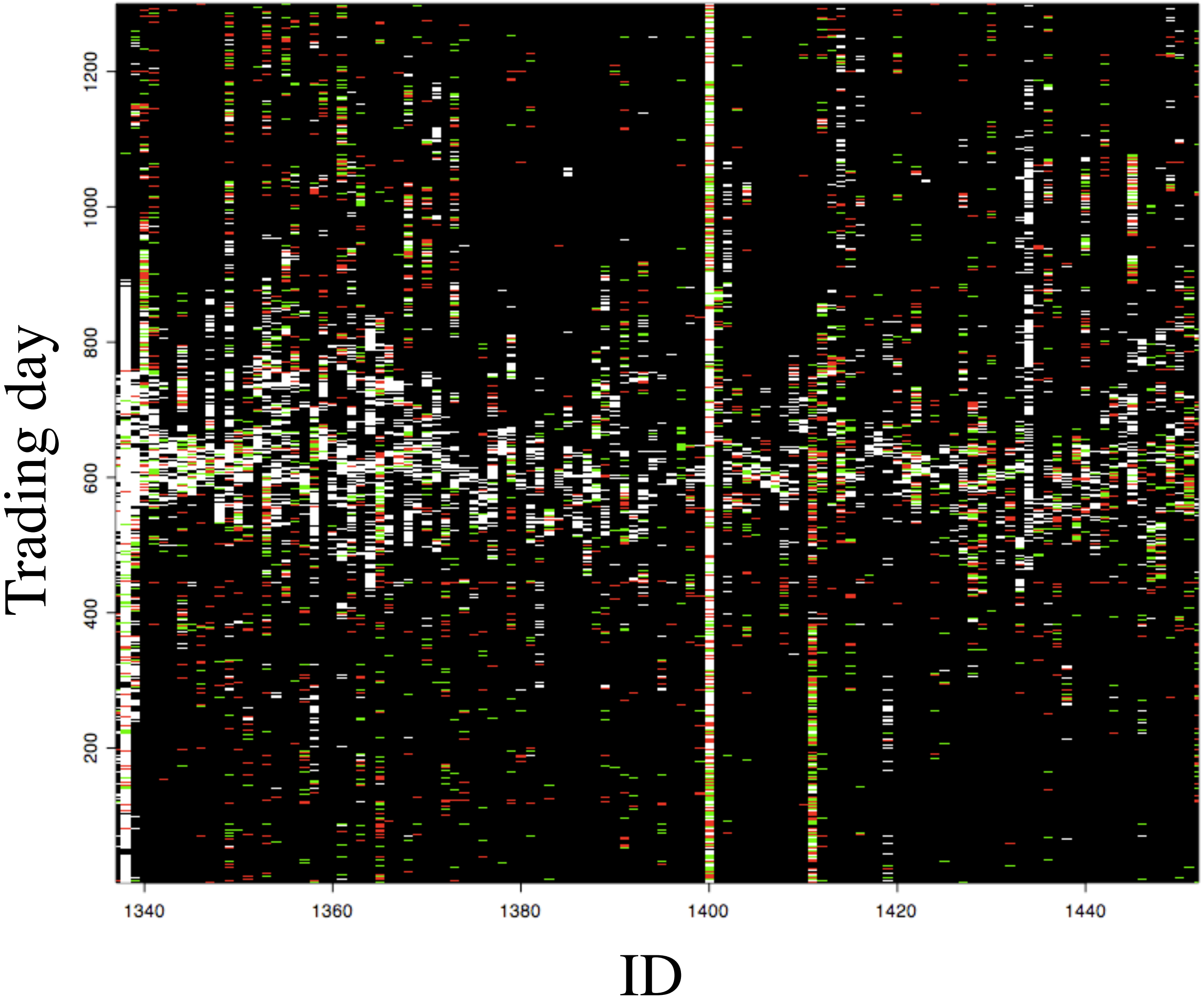


Over-expression of:

- Households investors;
- C1 (B.B) and C2 (S.S) links;
- Age cohort1941-1950
- Male gender

The trading activity is sometime pretty high over a period of time spanning a number of years

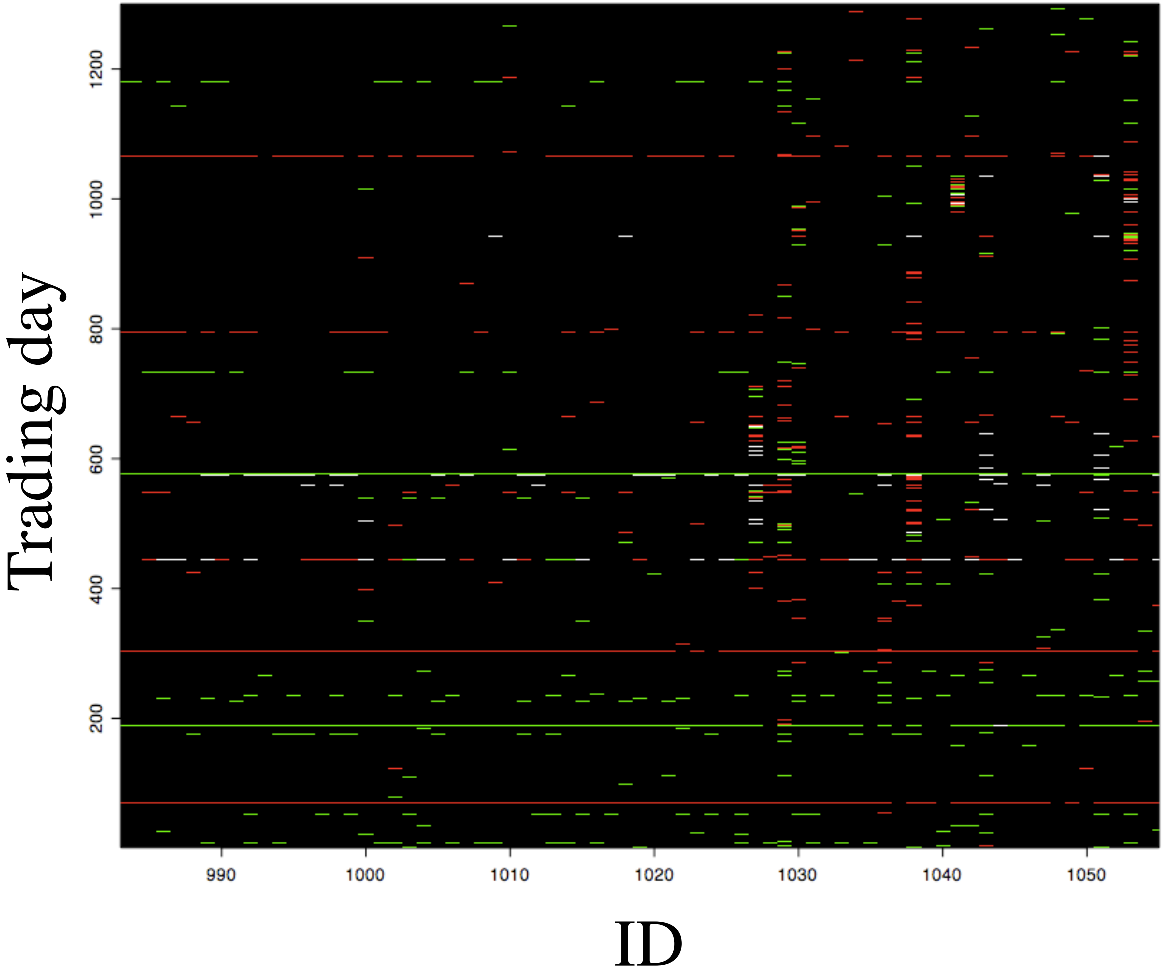
B4: 116 investors.



Over-expression of:
- C3 (BS.BS);

The trading activity is sometime pretty localized in a specific period of time.

B8: 73 investors.



Over-expression of:
- C2 (S.S);

In other cases the frequency of the trading activity is much lower but synchronicity in the trading decisions is seen on a time period spanning several years.

H = Households
C= Companies
G= Gov. Inst.
FI= Financial Inst.
NP= Non Profit Inst.
C1= B.B
C2= S.S
C3= BS.BS
C4= B.B S.S
C5= B.BS
C6= S.BS
C7= B.S
C8= B.B S.S BS.BS
C9= B.BS S.BS
0=Legal entity
1=1902-1940
2=1941-1950
3=1951-1960
4=1961-1970
5=1971-2000
0=Legal entity
1=Male
2=Female

Characterization of link and vertex attributes of clusters of the FDR network

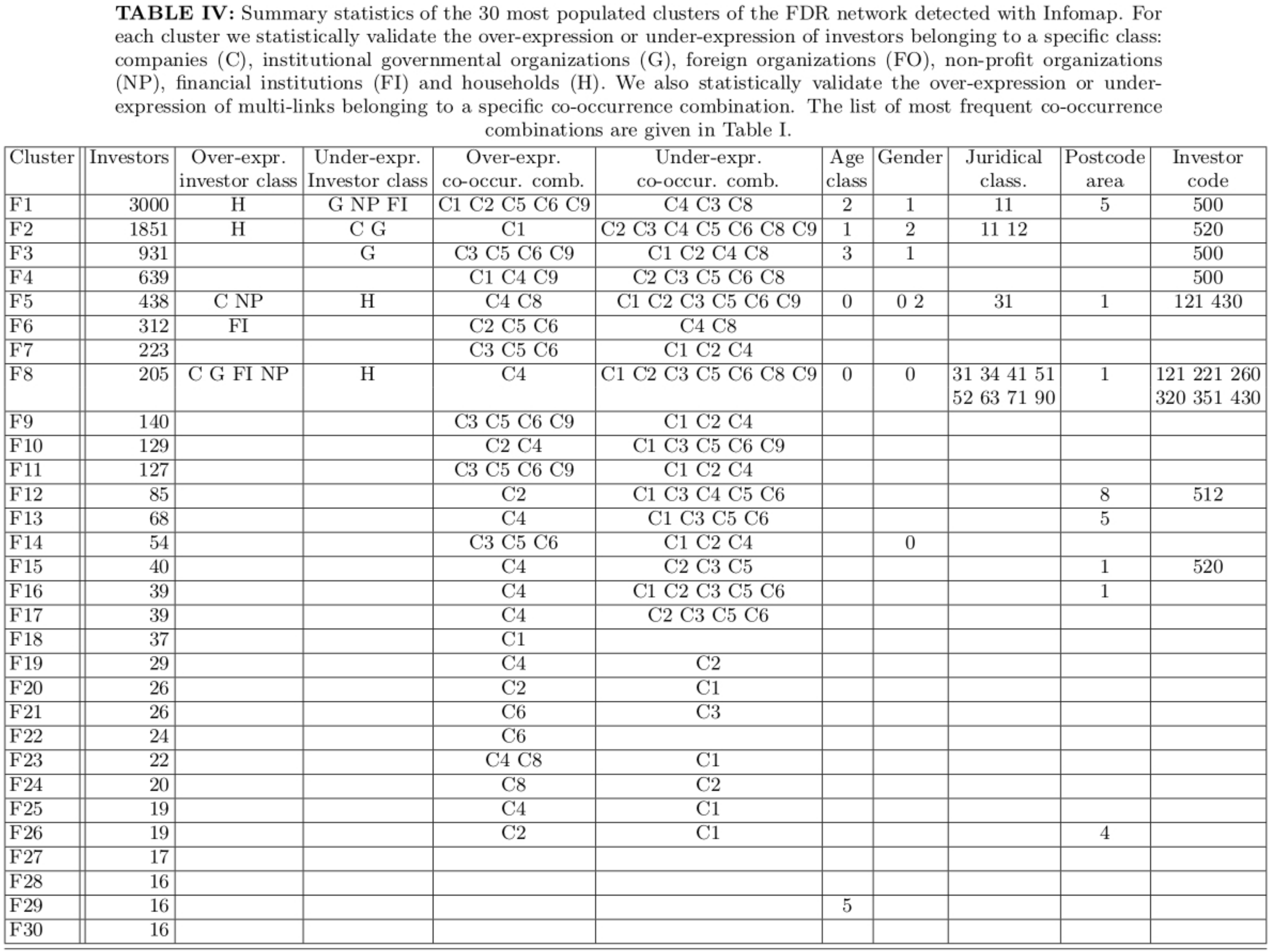
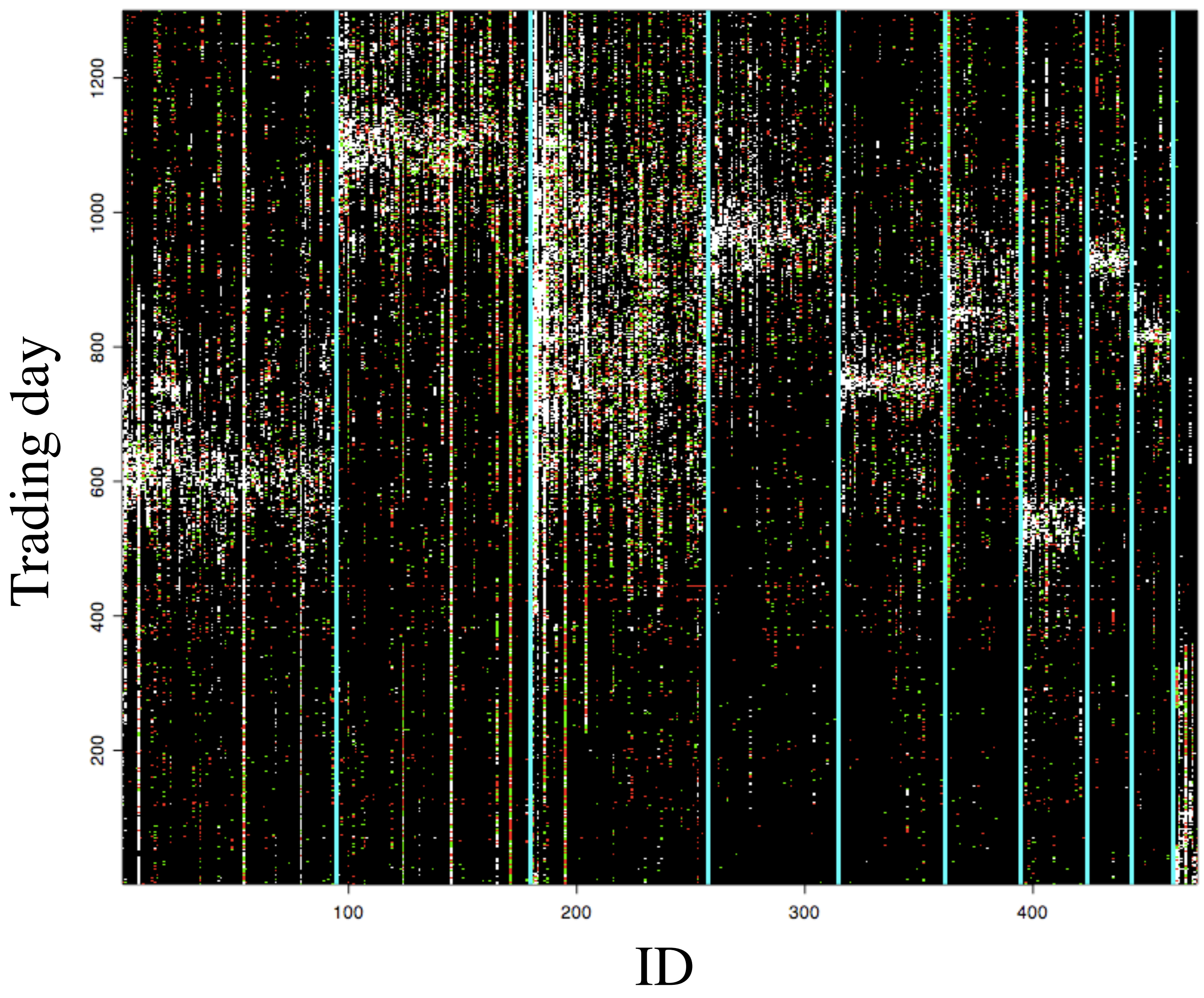


TABLE IV: Summary statistics of the 30 most populated clusters of the FDR network detected with Infomap. For each cluster we statistically validate the over-expression or under-expression of investors belonging to a specific class: companies (C), institutional governmental organizations (G), foreign organizations (FO), non-profit organizations (NP), financial institutions (FI) and households (H). We also statistically validate the over-expression or under-expression of multi-links belonging to a specific co-occurrence combination. The list of most frequent co-occurrence combinations are given in Table I.

Cluster	Investors	Over-expr. investor class	Under-expr. Investor class	Over-expr. co-occur. comb.	Under-expr. co-occur. comb.	Age class	Gender	Juridical class.	Postcode area	Investor code
F1	3000	H	G NP FI	C1 C2 C5 C6 C9	C4 C3 C8	2	1	11	5	500
F2	1851	H	C G	C1	C2 C3 C4 C5 C6 C8 C9	1	2	11 12		520
F3	931		G	C3 C5 C6 C9	C1 C2 C4 C8	3	1			500
F4	639			C1 C4 C9	C2 C3 C5 C6 C8					500
F5	438	C NP	H	C4 C8	C1 C2 C3 C5 C6 C9	0	0 2	31	1	121 430
F6	312	FI		C2 C5 C6	C4 C8					
F7	223			C3 C5 C6	C1 C2 C4					
F8	205	C G FI NP	H	C4	C1 C2 C3 C5 C6 C8 C9	0	0	31 34 41 51 52 63 71 90	1	121 221 260 320 351 430
F9	140			C3 C5 C6 C9	C1 C2 C4					
F10	129			C2 C4	C1 C3 C5 C6 C9					
F11	127			C3 C5 C6 C9	C1 C2 C4					
F12	85			C2	C1 C3 C4 C5 C6				8	512
F13	68			C4	C1 C3 C5 C6				5	
F14	54			C3 C5 C6	C1 C2 C4		0			
F15	40			C4	C2 C3 C5				1	520
F16	39			C4	C1 C2 C3 C5 C6				1	
F17	39			C4	C2 C3 C5 C6					
F18	37			C1						
F19	29			C4	C2					
F20	26			C2	C1					
F21	26			C6	C3					
F22	24			C6						
F23	22			C4 C8	C1					
F24	20			C8	C2					
F25	19			C4	C1					
F26	19			C2	C1				4	
F27	17									
F28	16									
F29	16					5				
F30	16									

The array representation of the trading activity



Conclusions

- Statistically validated networks can be obtained in event network occurring in a bipartite network;
- Over-expressed and under-expressed (with respect to a null hypothesis) links are highly informative;
- Familywise error needs to be taken into account to avoid false positive in large networks;
- Backbone of a network and negative unexpressed links can be detected;
- Statistically validated networks are useful for pre-processing and pruning noisy datasets;
- Discrete decisions can be detected in heterogeneous systems;
- Core of communities can be detected in incomplete and/or noise datasets;
- Communities can be characterized with respect to node attributes.
- Surveys can be investigated with this approach.

Our research group in Palermo, Italy
Observatory of Complex Systems

can open a one year research position ("assegno di ricerca") on networks in a socio-technical system (the air transportation system).

If you are interested
please send an email to

salvatore.micciche@unipa.it
or
rn.mantegna@gmail.com