istituto dei sistemi complessi

Consiglio Nazionale delle Ricerche

Bootstrapping Topological Properties and Systemic Risk of Complex Networks Using the Fitness Model.

Andrea Gabrielli (ISC – CNR, Rome, Italy)

Email: andrea.gabrielli@roma1.infn.it

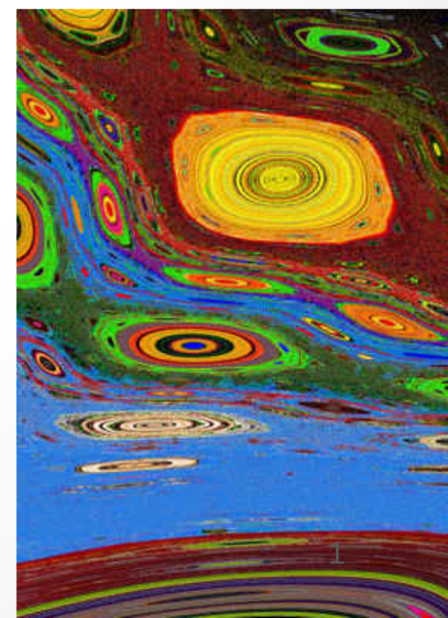
Collaborators:

Stefano Battiston (ETH, Zurich, CH)

Guido Caldarelli (IMT, Lucca, Italy)

Nicolò Musmeci (King’s College, London, UK)

Michelangelo Puliga (EHT, Zurich, CH)



Samarkand, UZ, 20-05-2013

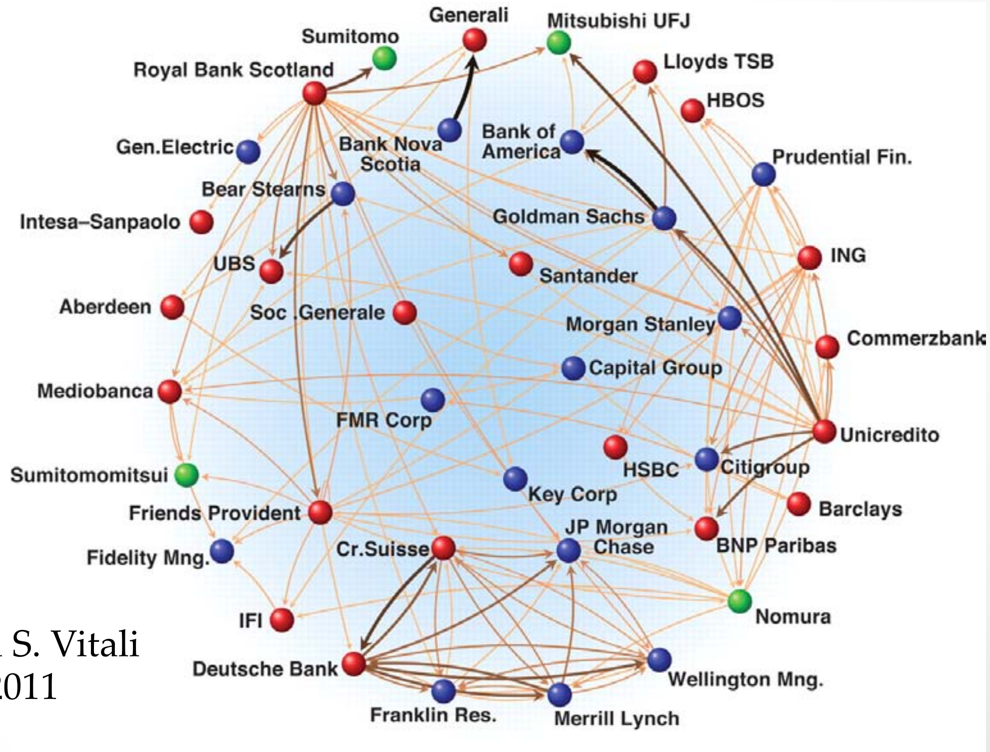
Summary

- We present a novel method to reconstruct global topological properties of a complex network from limited information.
- We assume to know for all nodes a non-topological quantity, called *fitness* correlated to the degree, and the degree only for a subset of the nodes.
- We then use a *fitness model*, calibrated on the subset of nodes for which degrees are known, in order to generate ensembles of networks to estimate properties of the original sample.
- We focus on global topological properties that are relevant for processes of contagion and distress propagation in networks: network density and k-core structure.
- We also study how well the resilience to distress propagation in the network can be estimated using our method: DebtRank
- We perform test on Exponential Random Graph model and on empirical networks taken from economic and financial contexts

The reconstruction of statistical properties of a network, when only partial information is available, is one of the outstanding and unresolved problems in the statistical physics of networks

Example: nodes = financial institutions and edges = various types of financial ties such as loans or derivative contracts

These ties, whose information is usually limited, result in dependencies among institutions and constitute the ground for the propagation of financial distress across the network



The Network of Global Corporate Control S. Vitali J. B. Glattfelder, S. Battiston, PLOS ONE, 2011

Status of the art: Maximal Entropy (ME) algorithms

The analysis of the systemic risk in partially unknown networks is performed in the Maximal Entropy scheme.

Standard methods assume that the network is fully connected.

The weights of the links are then obtained via a "maximum homogeneity" principle as each node is assumed to bear a similar level of dependence from all other nodes: full connection.

The method proceeds by looking for the weighted adjacency matrix that "minimizes the distance" from the uniform matrix while satisfying certain constraints (i.e. only certain "big" elements).

Such a matrix is found using the Kullback-Leibler divergence (i.e. –[entropy]) as the objective function to minimize

Problems of ME algorithms

- The hypothesis that the network is fully connected is a strong limitation of the ME algorithm.
- Empirical networks in economy, finance and social science often show instead a largely heterogeneous degree distribution.
- Such "dense reconstruction" leads in general to an underestimate of the systemic risk

In (1) a "sparse reconstruction" algorithm in which the Kullback-Leibler divergence is minimized with and arbitrary level $0<l<1$ of heterogeneity (i.e. of connections) has been proposed.

It is a more reliable algorithm, but suffers of a problem: heterogeneity is fixed a-priori.
What value of heterogeneity would be appropriate to choose?

(1) Mastromatteo, I., Zarinelli, E., Marsili, M.:. J. Stat. Mech. Theory Exp. **2012(03),** P03011 (2012)

Bootstrapping Method

N. Musmeci, S. Battiston, G. Caldarelli, M. Puliga, A. Gabrielli, J. of Stat. Phys., **151**, 220 (2013)
G. Caldarelli, A. Chessa, A. Gabrielli, F. Pammolli, M. Puliga, Nature Phys., **9**, 125 (2013)

We introduce a new general method that we name Bootstrapping Method (BM) to estimate both the topological properties of a network and its resilience to distress propagation starting from limited information with no assumptions on the connectivity.

It is based on the Exponential Random Graph model and the related Fitness model, by adding the condition of partial information

We study how the accuracy of the estimation depends upon the size of the subset of nodes for which the information is available

To validate our method we use both synthetic networks as well as examples of real economic systems: World Trade Web (WTW), the e-mid interbank loan network

Exponential Random Graph Model (ERGM)

J. Park and M.E.J. Newman, Phys. Rev. E, **70**, 066117 (2004)

Let us consider binary undirected graphs with fixed N nodes

$\{C_a\}$ = set of graph properties that we want to fix *in some way*

ERG defines the maximally random ensemble Ω of graphs with N nodes compatible with the constraints (canonical ensemble):

$$\langle C_a \rangle \equiv \sum_G C_a(G) P(G) = C_a^* \quad \forall a \quad (1)$$

G = generic graph of the ensemble with N nodes

$P(G)$ = measure on the ensemble $\Rightarrow$ it is found by maximizing the entropy

$$S(G) = -\sum_G P(G) \log(G) \text{ with the constraints (1)} \Rightarrow$$

$$P(G) = \frac{1}{Z} \exp[-H(G)]$$

$$H = \sum_a \theta_a C_a(G) \text{ with } \theta_a = \text{ Lagrange multipliers associated to (1)}$$

If $\{C_a\} = \{k_i\}$ $i=1,\dots,N$ ➔ $H(G)=\sum_i \theta_i k_i$

If we call $x_i=e^{-\theta_i}$ then the ensemble is simply defined by

$$p_{ij} = \frac{x_i x_j}{1+x_i x_j} = \text{prob. nodes } i \text{ and } j \text{ connected}$$

$$\langle k_i \rangle = \sum_{j(\neq i)}^{1,N} p_{ij}; \quad \langle k_i^{nn} \rangle = \frac{\sum_{j(\neq i)}^{1,N} \sum_{k(\neq j)}^{1,N} p_{ij} p_{jk}}{\langle k_i \rangle}; \quad \langle C_i \rangle = \frac{\sum_{j(\neq i)}^{1,N} \sum_{k(\neq i,j)}^{1,N} p_{ij} p_{jk} p_{ki}}{\langle k_i \rangle [\langle k_i \rangle - 1]}$$

Fixing $\{x_i\}$, by some intuition, is the same of fixing $\{\langle k_i \rangle\}$

In particular for small x ➔ $\langle k_i \rangle \approx \sum_j x_i x_j \sim x_i$

Fitness model

D. Garlaschelli and M. Loffredo, Phys. Rev. Lett., **93**, 188701 (2004)

G. De Masi, G.Iori, G. Caldarelli: Phys. Rev. E A **74**, 066112 (2006)

Let us suppose to have incomplete information about the topology of a given real network G_0 of N nodes

We assume to know:

1) the degree $k_i=k_i^0$ for a subset of nodes I of the network G_0;
2) a non-topological property y_i, called fitness, assumed to be *roughly linearly correlated* to k_i for all nodes

This happens for instance for the binary undirected WTW where nodes are countries and y_i is the national GDP

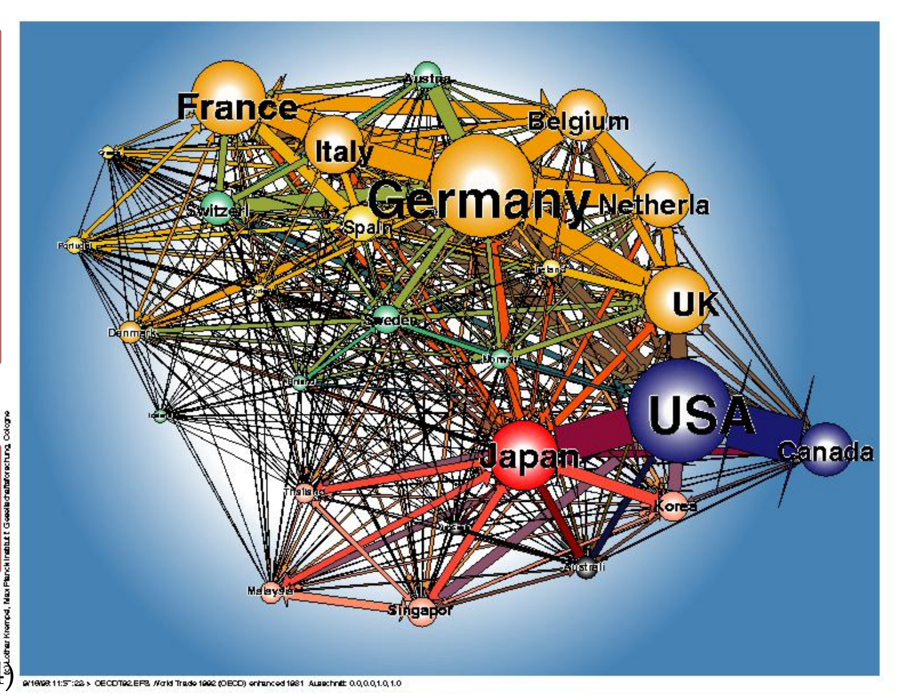
The World Trade Web (WTW)

The WTW is a weighted directed network defined by the exchange of wealth between countries (vertices)

Each vertex **i** is characterized by its total wealth (GDP) y_i

The undirected binary version of the WTW is defined by the graph in which the unweighted link between two countries **i** and **j** is present if a non zero flow of wealth exists in any direction between them

It is statistically well reproduced by a ERGM with $x_i = a y_i$ with $a \approx 10^{-2}$



D. Garlaschelli and M. Loffredo, Phys. Rev. Lett., **93**, 188701 (2004)

Bootstrapping method

Having only the above partial information, to estimate the statistical features of some property $s(G_0)$ of the real network G_0, we impose the maximal entropy compatible with the constraints

↓

The network G_0 is assumed to be extracted from a suitable ensemble of ERG including the available information: $\{y_i\}_N$, $\{k_i^0\}_I$

Maximal likelihood

Each known value of the non-topological property y_i is assumed to be proportional to the fitness, denoted as x_i : $\sqrt{z y_i} = x_i$

Therefore $$p_{ij} = \frac{z y_i y_j}{1 + z y_i y_j} = \text{prob. nodes } i \text{ and } j \text{ connected}$$

How to determine the unknown parameter z?

Case 1: |I|=N (complete information)
Assuming small fluctuations in the canonical ensemble

$$\langle L \rangle \equiv \frac{1}{2}\sum_{i=1}^{N}\langle k_i \rangle \equiv \frac{1}{2}\sum_{i=1}^{N}\sum_{j(\neq i)}^{1,N} p_{ij} = L_0 = \frac{1}{2}\sum_{i=1}^{N} k_i^0 \rightarrow \text{Eq. for } z$$

D. Garlaschelli and M. Loffredo, Phys. Rev. Lett., **93**, 188701 (2004)

This has been used for WTW, the network of equity investments in the stock market, the interbank market

Case 2: |I|<N (partial information, typical for financial networks)
N. Musmeci, S. Battiston, G. Caldarelli, M. Puliga, A. Gabrielli, J. of Stat. Phys., **151**, 220 (2013)

$$\sum_{i\in I}\langle k_i \rangle \equiv \sum_{i\in I}\sum_{j(\neq i)}^{1,N} p_{ij} = \sum_{i\in I} k_i^0 \rightarrow \text{Eq. for } z$$

We study how accuracy increases with |I|

Having the estimate of z, the ERG ensemble is completely defined and all the averages of the properties of G_0 can be estimated

Test of BM on synthetic networks

We build a ERG ensemble in the following way:
N=185 (as WTW in year 2000)
Fix y_i=GDP in WTW
Use $z=z_0=10^4$ (fitness model in WTW with complete information)
$p_{ij}= zy_iy_j/(1+zy_iy_j)$ (note we do not impose k_i)

In this way we can evaluate all the ensemble properties of this ERG

We focus on three important quantities for the contagion/distress propagation in networks:
1) density of links D = # of links divided for N(N-1)/2;
2) degree of the main core, k^{main} (the k-core is the largest connected subgraph with all nodes with degree ≥k);
3) size of the main core, S^{main} (i.e., # of nodes).

For all these properties we evaluate averages $\langle D\rangle_0$, $\langle k_{main}\rangle_0$, $\langle S_{main}\rangle_0$

Estimate of the ensemble properties with partial information

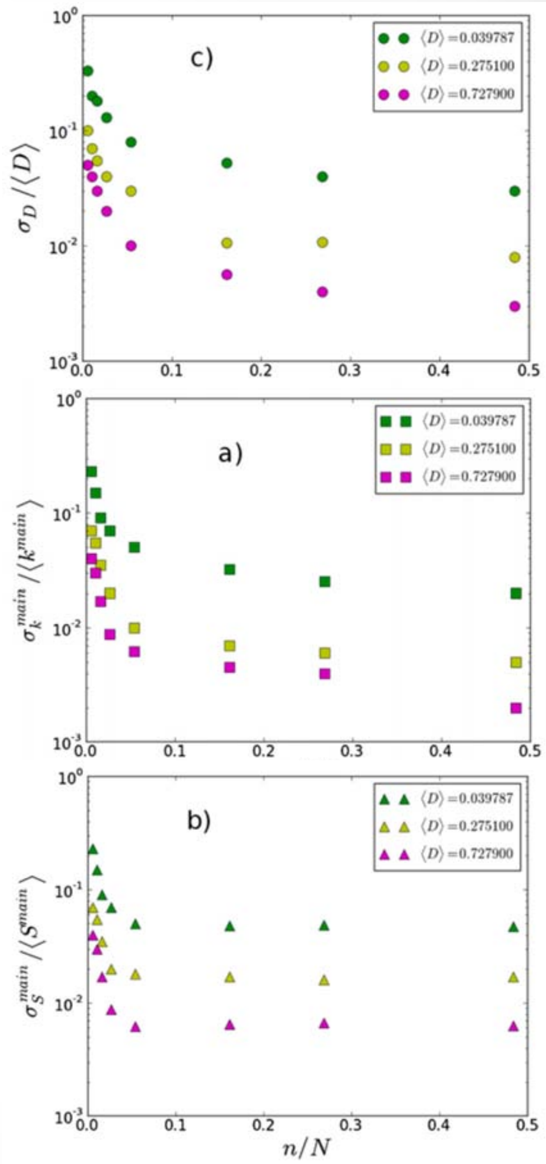
Let us consider a generic subset of nodes such that $|I|=n<N$
We consider to know y_i for all N nodes and k_i only for I.

As seen, we can estimate z from above eqs.

$$\sum_{i\in I}\sum_{j(\neq i)}^{1,N} p_{ij} = \sum_{i\in I} k_i^0 \rightarrow z'$$

With the estimate z' of z_0, we can build another ERG ensemble and evaluate the averages $\langle D\rangle_I$, $\langle k_{main}\rangle_I$ and $\langle S_{main}\rangle_I$ and to study their deviation from $\langle D\rangle_0$, $\langle k_{main}\rangle_0$ and $\langle S_{main}\rangle_0$ at varying n from 1 to N

Already at $n/N<0.1$ we have a good estimate



Test on real networks (1): World Trade Web

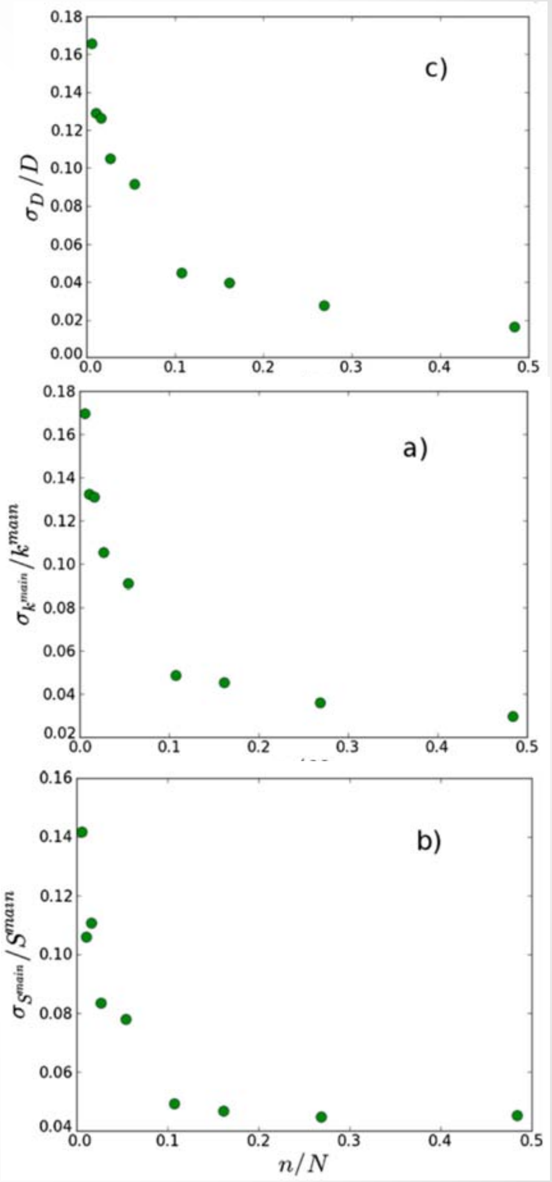
In this test we compare the BM with the WTW binary network (N=185)

For this network we evaluate the exact values of D, k_{main} and S_{main}

We take an arbitrary subset I of n<N nodes and build a ERG ensemble with N nodes with $\{y_i\}_N$ and $\{k_i\}_I$ evaluating z' as above

We study the accuracy of $\langle D \rangle_I$, $\langle k_{main} \rangle_I$ and $\langle S_{main} \rangle_I$ at varying n from 1 to N

Again we get a good approximation already for n/N<0.1



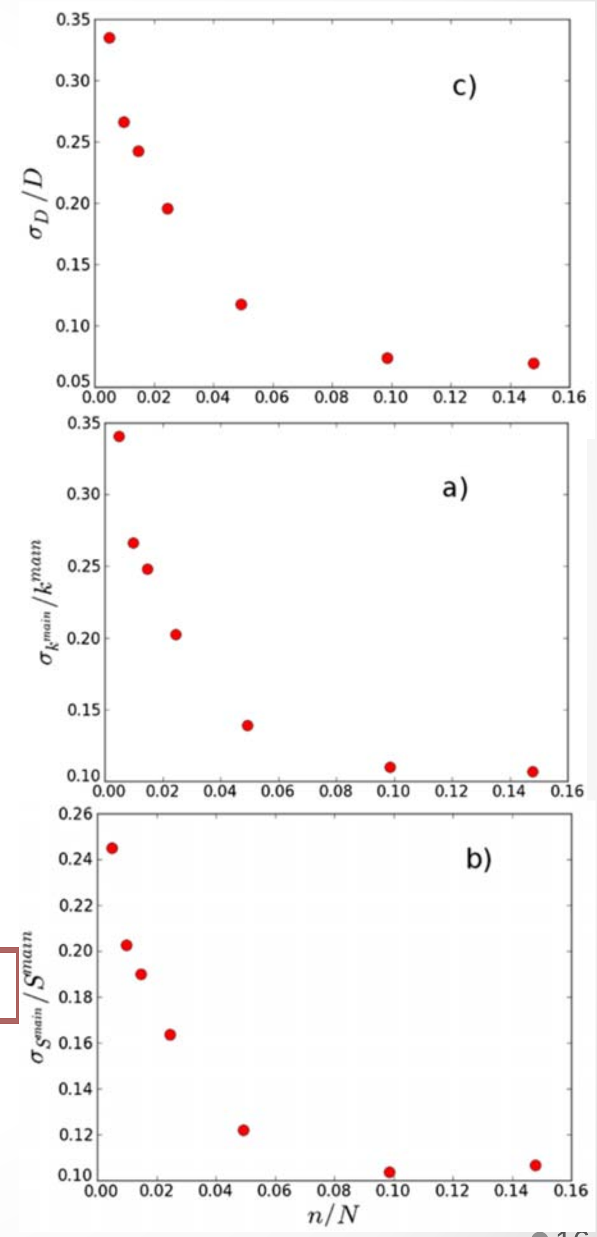
Test on real networks (2): E-mid network

E-mid network=interbank loan network

G. De Masi, G.Iori, G. Caldarelli: Phys. Rev. E A **74**, 066112 (2006)

We perform exactly the same analysis as for the WTW

Again good approximation already for n/N<0.1



Test of BM: DebtRank a Measure of Systemic Risk

To estimate the resilience of the networks we use DebtRank (DR)

S. Battiston, M. Puliga, R. Kaushik, P. Tasca, G. Caldarelli, Sci. Rep. **2**, 541 (2012)

We also use Group DebtRank (GDR), which is a measures of the impact of a small shock on all the nodes in the network, due to the reverberations across links in the network.

It is a recursive algorithm to compute the impact on the whole network of an initial shock ψ to one node (DR) and to all nodes (GDR) using the weights w_{ij} of the adjacency matrix

The rescaling factor $0<\alpha<1$ determines the scale of the impact along each link in the network (propagation).

For each real network (WTW and E-mid) our goal is to test how well DR and GDR are estimated by the network bootstrap method.

Until now binary matrix, instead now we work on weighted networks

How to assign unknown weights to links and knowing only those for the set I of n<N nodes?

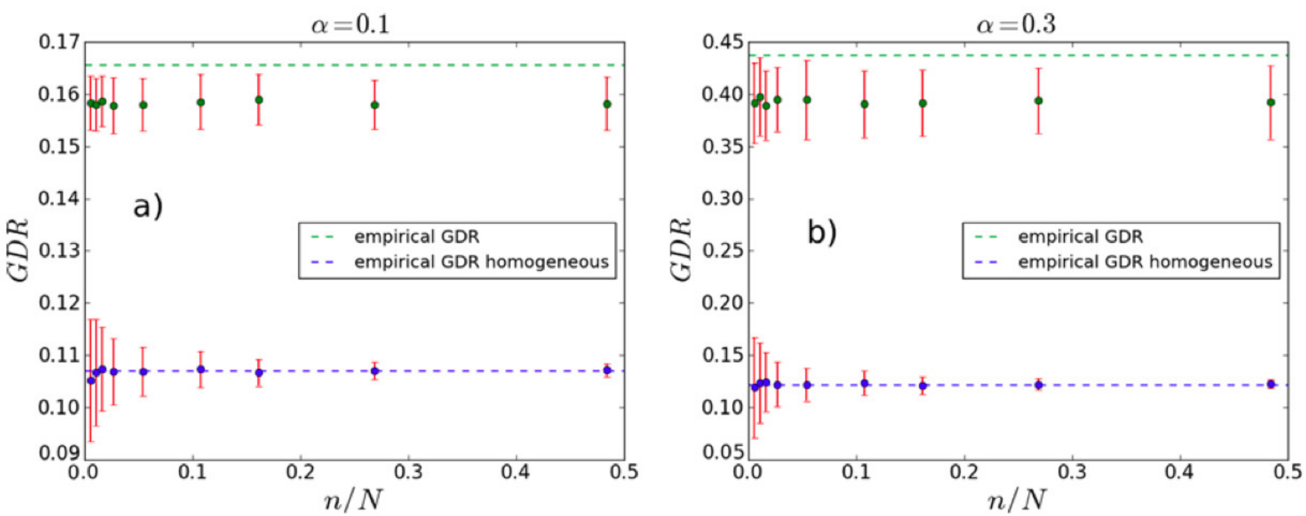
1) Homogeneous network: Compute the average weight by averaging the elements of the w_{ij} matrix associated to the n<N nodes. Use this value as *homogeneous* weight for all nodes;

2) Assign to each node a weight similarly to what done in a gravity model where the link l_{ij} has a weight proportional to the product of the GDPs *GDPi · GDPj* (for E-mid network GDP→total out-strength).

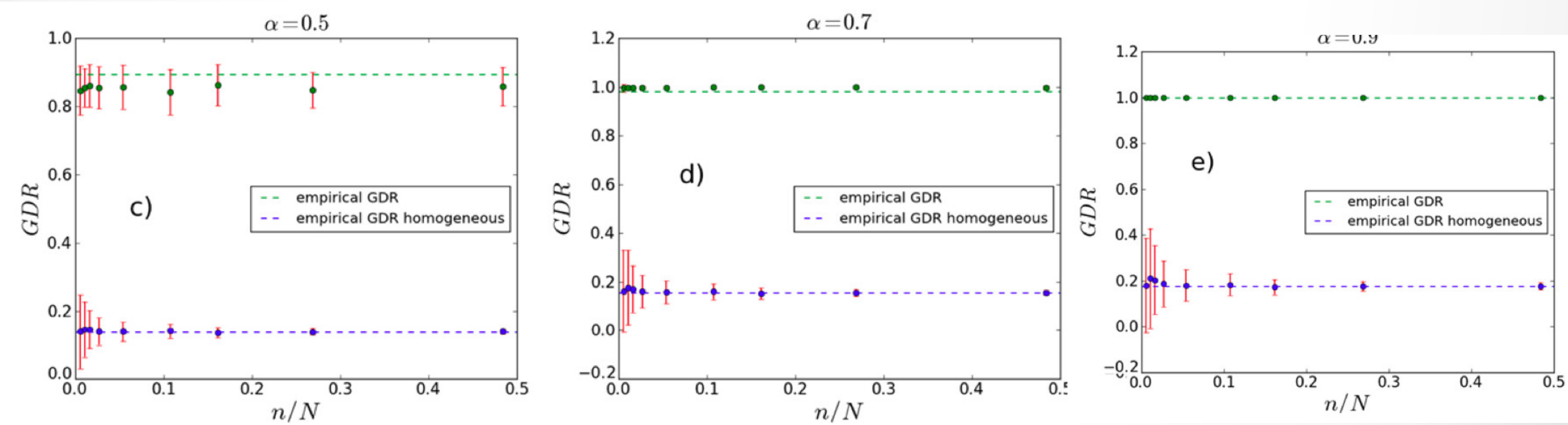
G. Fagiolo, J. Econ. Interact. Coord. **5(1),** 1–25 (2010)

3) Directed network is imposed symmetrically and weights appropriately renormalized.

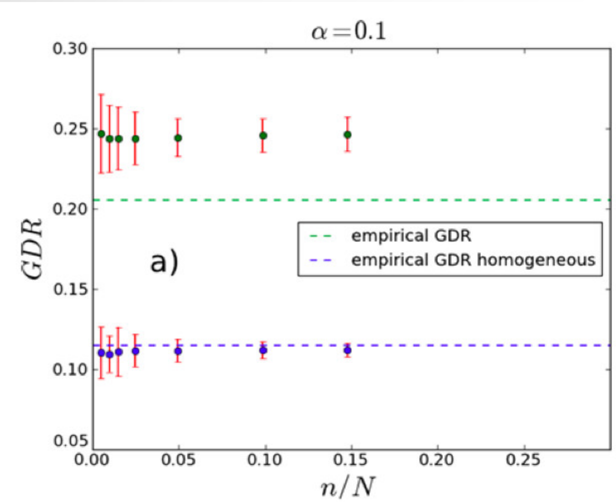
GDR on WTW

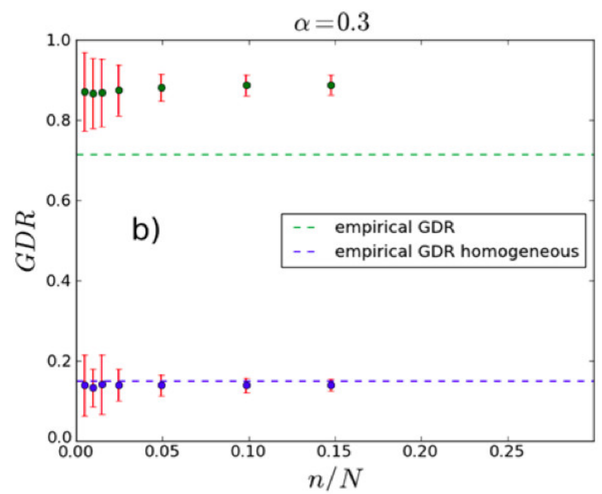


Blue: homogeneous
Green: gravity model
vs empirical

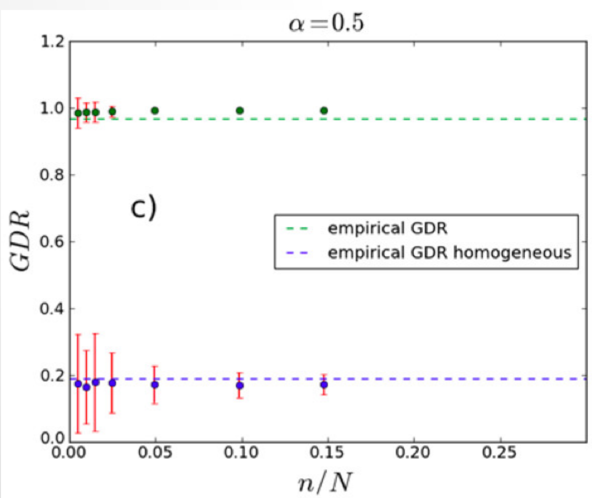


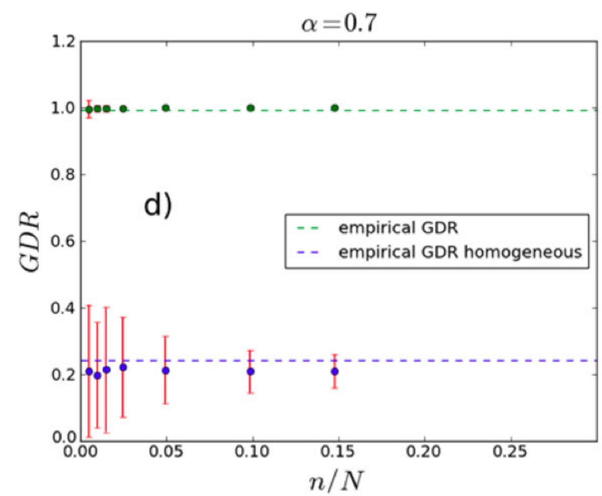
GDR for E-mid interbank network





Blue: homogeneous
Green: gravity model
vs empirical





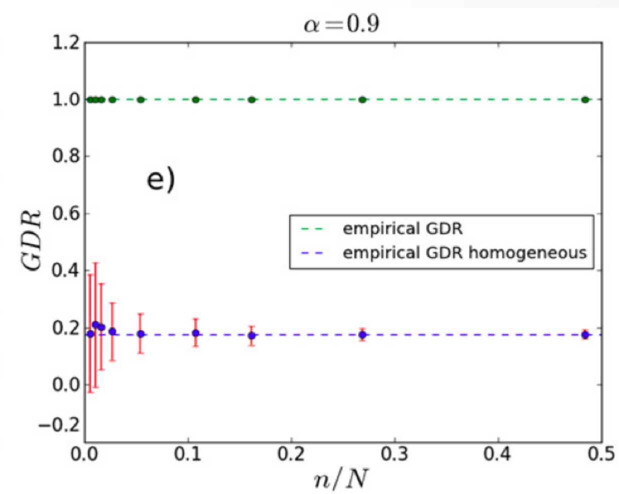


Table 1 The values of DebtRank and the GDP rank (year 2000) for the 20 biggest countries in the WTW network. Notice that the ranking according to DebtRank agrees only in part with the ranking by the GDP. Depending on the size of the exports volume, each country can be more or less affected by a shock on the others countries. The values are obtained setting $\alpha = 0.5$ and $\psi = 1$

Country	DebtRank	GDP rank (2000)
USA	0.48	1
JPN	0.32	2
CAN	0.26	8
CHN	0.23	6
DEU	0.23	3
MEX	0.18	10
GBR	0.17	4
FRA	0.16	5
ITA	0.12	7
NLD	0.10	15
KOR	0.09	12
TWN	0.09	16
BEL	0.08	20
ESP	0.08	11
SGP	0.07	39
MYS	0.05	40
CHE	0.05	18
BRA	0.05	9
IRL	0.04	38
AUS	0.04	14

Conclusions

- Recovering the statistical properties of the topology of a network from partial information is a fundamental problem in many fields (e.g. finance, epidemiology etc.)

- Its main application is in the forecast of the effects of distress propagation in the network

- Bootstrapping method, based on fitness model, is able in important cases to reconstruct most of the main properties of the network from limited topological information and the knowledge of non-topological information on all nodes ~linearly correlated to connectivity

- This can be the base to predict from such partial information the effect of the propagation epidemy: e.g. using the DebtRank algorithm

- Extensions: weighted and directed networks, non-linear correlations between fitnesses and topological properties.