Networks, proximity based networks and statistically validated networks in Finance

Rosario Nunzio Mantegna
Center for Network Science, Central European University, Hungary
Department of Physics and Chemistry, University of Palermo, Italy

Overview

- Networks and proximity based networks;
- Proximity based networks in finance and economics;
- Other types of "association networks";
- Statistically validated networks in credit market and in stock markets;
- Conclusions.

Networks in complex systems

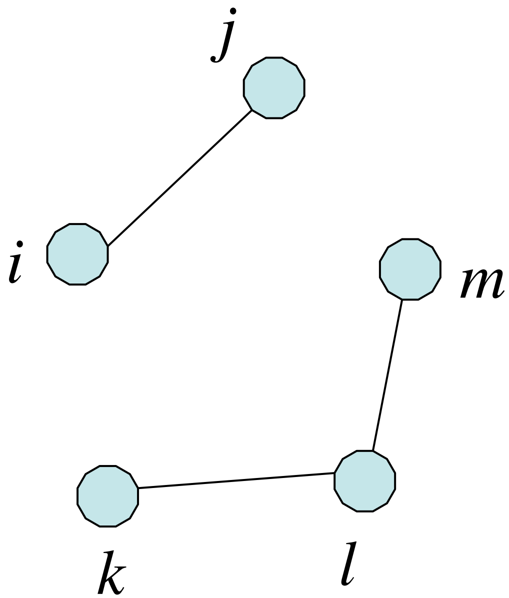
In several systems a link between two nodes in a network is directly observable. Examples are: links between cities due to the presence of at least one flight, links between routers of the Internet, links between WWW pages, etc.

There are other systems where a link is set between two nodes which are presenting a given amount of similarity. Examples are: Similarity of asset returns time series, similarity of exported products, similarity of portfolio's allocation, etc.

The networks obtained starting from a proximity (similarity or dissimilarity) measure are called "similarity-based networks" or "association networks".

Two specific examples:

Event or relation defined networks



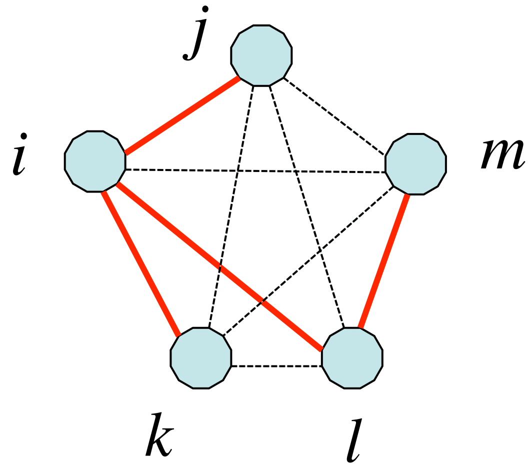
Example:
nodes are banks
links are credit relationships

Proximity-based networks

Example: 1) Consider
Portfolio of bank i
portfolio of bank j
.....
portfolio of bank m

2) Estimate similarity/distance between each pair of banks;

3) Extract a weighted network from a similarity/distance matrix.



Generally speaking, let us consider a set of vertices V. Each vertex $v \in V$ is characterized by a vector of attributes $x \in R^n$

By using the attributes it is possible to estimate a user-specified proximity measure s_{ij} or d_{ij} for each pair of vertices $i, j \in V$

In other words, similarity based networks or association networks are usually constructed by using a proximity measure which is not directly observable, but it is obtained from the set of attributes $\{x_i\}$.

Let $x \in R^n$ be the vector of continuous attributes that characterize the set of vertices V.

A standard similarity measure between vertex pairs is $s_{ij} = \rho_{ij}$ where

$$\rho_{ij} = cor\left(X_i, X_j\right) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

is the Pearson correlation coefficient.

In the presence of discrete attributes, other similarity measures such as, for example, the Jaccard coefficient can be considered.

By using the correlation coefficient as a similarity measure, a network can be obtained by considering that an edge is present between node i and node j when ρ_{ij} is different from zero (or it is above a certain threshold).

Under this protocol G (V, E) is characterized by the edge set defined as

$$E = \left\{ \{i, j\} \in V^{(2)} : \rho_{ij} \neq 0 \right\}$$

Such a protocol therefore needs many statistical tests of the hypothesis

$$H_0 : \rho_{ij} = 0 \quad \text{versus} \quad H_1 : \rho_{ij} \neq 0$$

By using the empirical observations of the *n* records of the x vector of attributes, the empirical correlation coefficient is estimated as

$$\hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}$$

Under the assumption (quite reasonable but not always verified in empirical data) that each pair of variables (x_i,x_j) has a bivariate Gaussian distribution, the density of $\hat{\rho}_{ij}$ under $H_0 : \rho_{ij} = 0$ is known but has not a simple analytical form.

To obtain p-values, an approximate approach based on Fisher's transformation is usually followed. In fact the variable

$$z_{ij} = \tanh^{-1}\left(\hat{\rho}_{ij}\right) = \frac{1}{2}\log\left[\frac{1+\hat{\rho}_{ij}}{1-\hat{\rho}_{ij}}\right]$$

has a density function that is well approximated by a Gaussian density with zero mean and variance $1/(n-3)$ under the hypothesis of bivariate Gaussian distribution. The approximation is asymptotic in n but the convergence is quite accurate already for moderate values of n.

The estimation of a similarity based network can then proceed in terms of the following steps:

(i) estimation of $\hat{\rho}_{ij}$ for the $N_v(N_v$-1)/2 distinct correlation coefficients;

(ii) evaluation of the *p*-values associated with each test assuming $H_0 : \rho_{ij} = 0$. An assumption about the statistical properties of the bivariate stochastic process is needed at this step;

(iii) multiple hypothesis test correction.

The multiple hypothesis test correction can be done as a Bonferroni correction or by using the False Discovery Rate control proposed by Benjamini and Hochberg¶

Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57, 289-300 (1995).

When the estimation of the correlation coefficient is accurate (e.g. records of the attributes well described in terms of Gaussian random variables and estimation performed on a large number of records) most (and sometimes all) of the correlation coefficients are rejecting the hypothesis $H_0 : \rho_{ij} = 0$.

This implies that the similarity based networks obtained with this protocol can be a graph which is often overlapping or very close to a fully connected network.

It is therefore worth devising protocols able to filter the most important information present in similarity based networks.

Filtering based on hierarchical clustering: the single linkage.

By starting from a correlation matrix (which is a similarity measure)

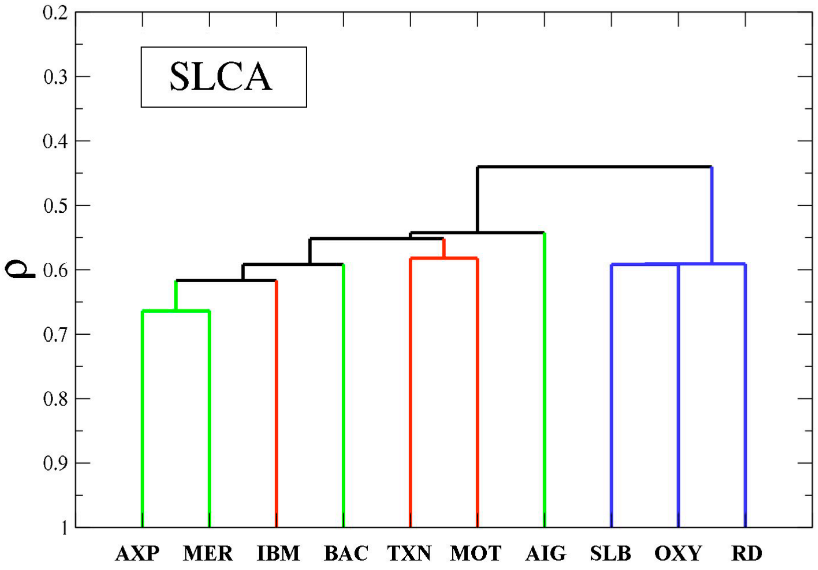
	AIG	IBM	BAC	AXP	MER	TXN	SLB	MOT	RD	OXY
AIG	1	0.413	0.518	0.543	0.529	0.341	0.271	0.231	0.412	0.294
IBM		1	0.471	0.537	0.617	0.552	0.298	0.475	0.373	0.270
BAC			1	0.547	0.591	0.400	0.258	0.349	0.370	0.276
AXP				1	0.664	0.422	0.347	0.351	0.414	0.269
MER					1	0.533	0.344	0.462	0.440	0.318
TXN						1	0.305	0.582	0.355	0.245
SLB							1	0.193	0.533	0.592
MOT								1	0.258	0.166
RD									1	0.590
OXY										1

AXP	MER	0.664
IBM	MER	0.617
SLB	OXY	0.592
BAC	MER	0.591
RD	OXY	0.590
TXN	MOT	0.582
IBM	TXN	0.552
AXP	BAC	0.547
AIG	AXP	0.543
AXP	IBM	0.537
SLB	RD	0.533
MER	TXN	0.533
AIG	MER	0.529
AIG	BAC	0.518
IBM	MOT	0.475
MOT	MER	0.462
MER	RD	0.440
AXP	TXN	0.422

…….

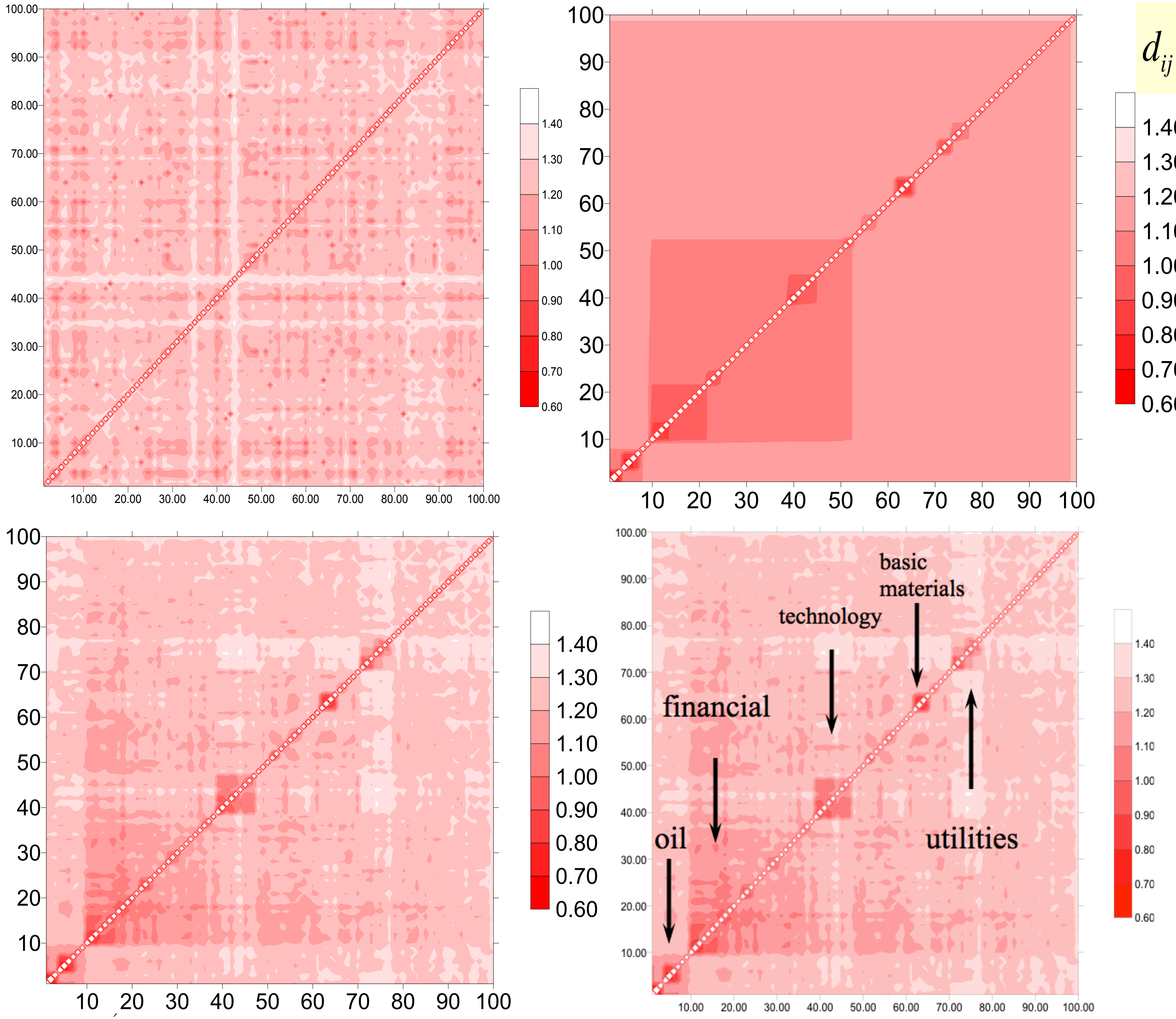
By applying the single linkage clustering procedure one obtains a hierarchical tree that has associated a (ultrametric) simplified matrix having only n-1 distinct correlation coefficients.

	AIG	IBM	BAC	AXP	MER	TXN	SLB	MOT	RD	OXY
AIG	1	0.543	0.543	0.543	0.543	0.543	0.440	0.543	0.440	0.440
IBM		1	0.591	0.617	0.617	0.552	0.440	0.552	0.440	0.440
BAC			1	0.591	0.591	0.552	0.440	0.552	0.440	0.440
AXP				1	0.664	0.552	0.440	0.552	0.440	0.440
MER					1	0.552	0.440	0.552	0.440	0.440
TXN						1	0.440	0.582	0.440	0.440
SLB							1	0.440	0.590	0.592
MOT								1	0.440	0.440
RD									1	0.590
OXY										1



$C^{<}_{SL}$

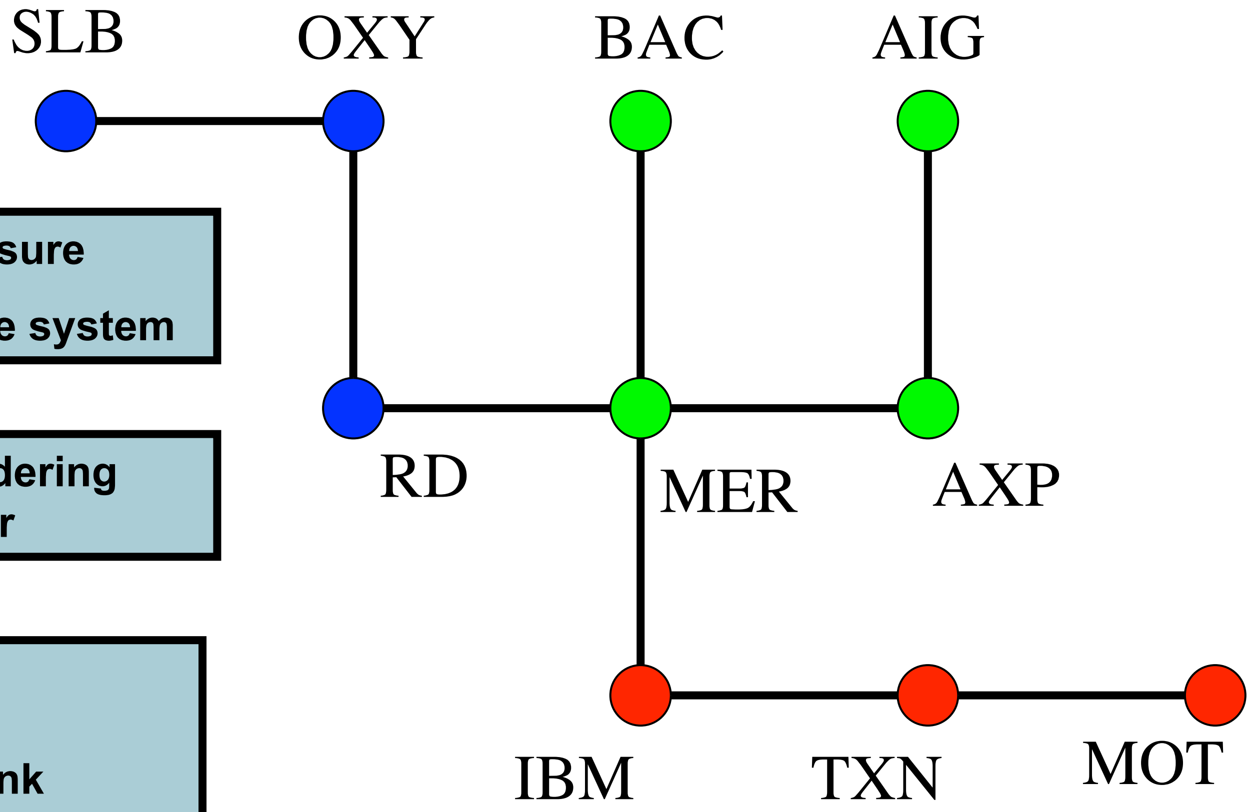
Filtering selects part of the information of a proximity matrix



$$d_{ij} = \sqrt{2\left(1-\rho_{ij}\right)}$$

100 highly capitalized stocks of US equity markets 1995-1998 daily data

Minimum Spanning Tree



Define a **proximity measure**
between the elements of the system

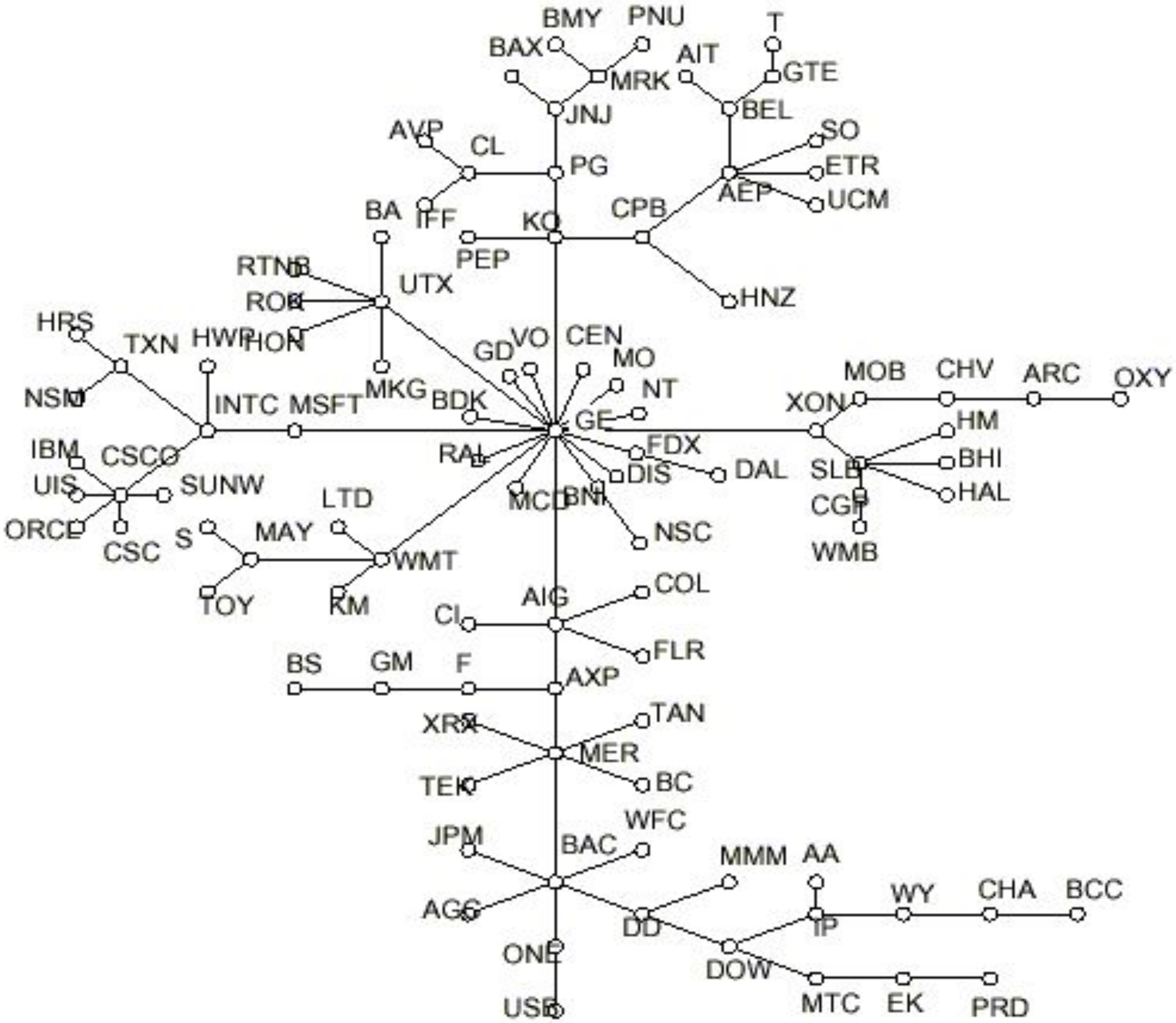
Construct the list ***P* by ordering proximities in order**

Starting from the first element of *P*,
add the corresponding **link**
if and only if
the graph is still a Forest or a Tree

Minimum Spannig Tree
MST

Correlation based trees and hierarchical trees do NOT carry the same amount of information.

Minimum Spanning Tree (MST)



N=100 (NYSE)
daily returns
1995-1998
T=1011

R. N. Mantegna, EPJ B 11, 193 (1999) G. Bonanno et al., Quant. Fin. 1, 96 (2001)

How to construct a graph richer of links but including the MST?

The Planar Maximally Filtered Graph (PMFG) is

- a topologically planar graph;
- connecting all nodes of the graph by keeping the shortest links and allowing at least 3 links for each node;
- topologically embedded in a surface of genus 0;
- a graph allowing loops.

The genus of a graph is the minimum number of handles that must be added to a surface to embed the graph without any crossings.
A planar graph therefore has graph genus 0.

MST and Planar Maximally Filtered Graph (PMFG)

Define a **similarity measure between the elements of the system**

Construct the list ***S*** **by ordering similarities in decreasing order**

Starting from the first element of *S*, **add** the corresponding **link** **if** and only if **the graph is still a Forest or a Tree**

Minimum Spannig Tree
MST

Starting from the first element of *S*, **add** the corresponding **link** **if** and only if **the graph is still Planar (g=0)**

Planar Maximally Filtered Graph
PMFG

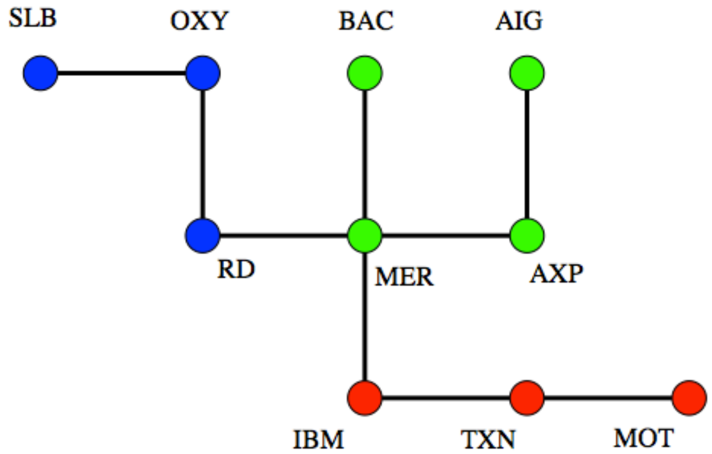
M. Tumminello et al, PNAS USA 102, 10421 (2005)

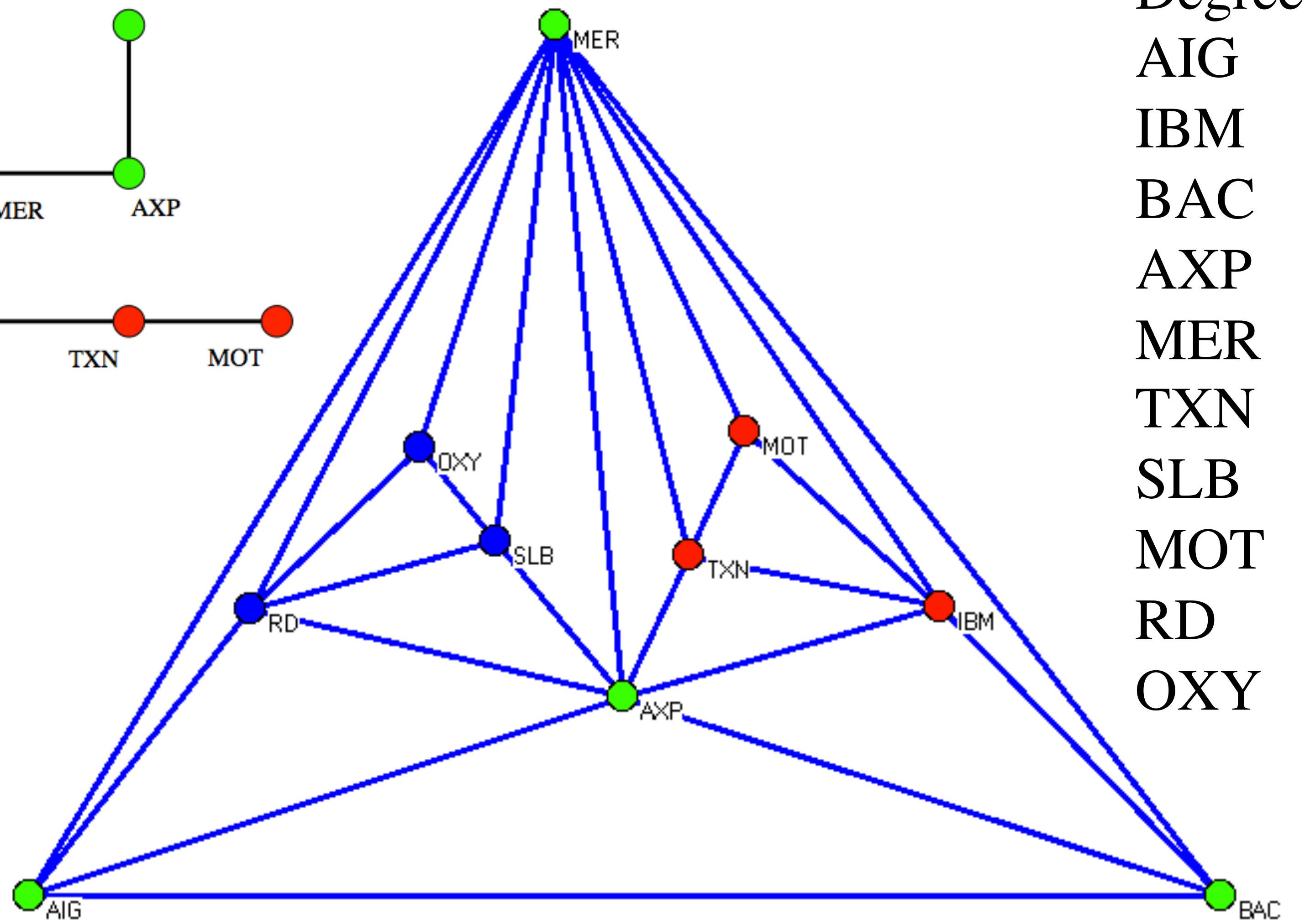
MST and agglomerative planar graphs

The Minimum Spanning Tree is always included into the Planar Maximally Filtered Graph or in any graph embedded in a surface of genus *g* and selected with a constructing algorithm similar to the one used for MST and PMFG.

The hierarchical properties of the graphs obtained with this constructing algorithm are the same as the one of the MST (they are characterized by the same clusters).

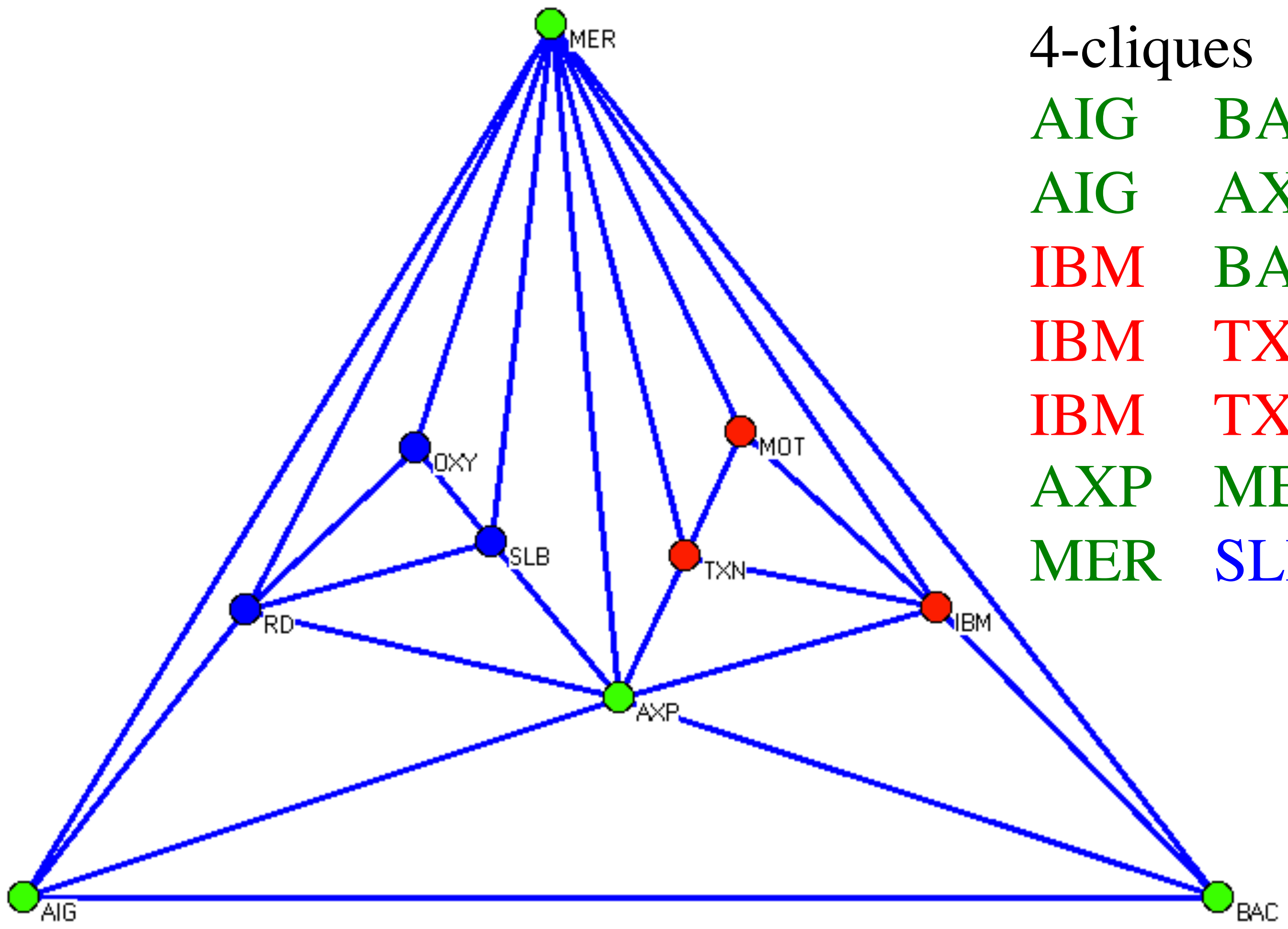
The example with the 10 stocks





Degree	
AIG	4
IBM	5
BAC	4
AXP	7
MER	9
TXN	4
SLB	4
MOT	3
RD	5
OXY	3

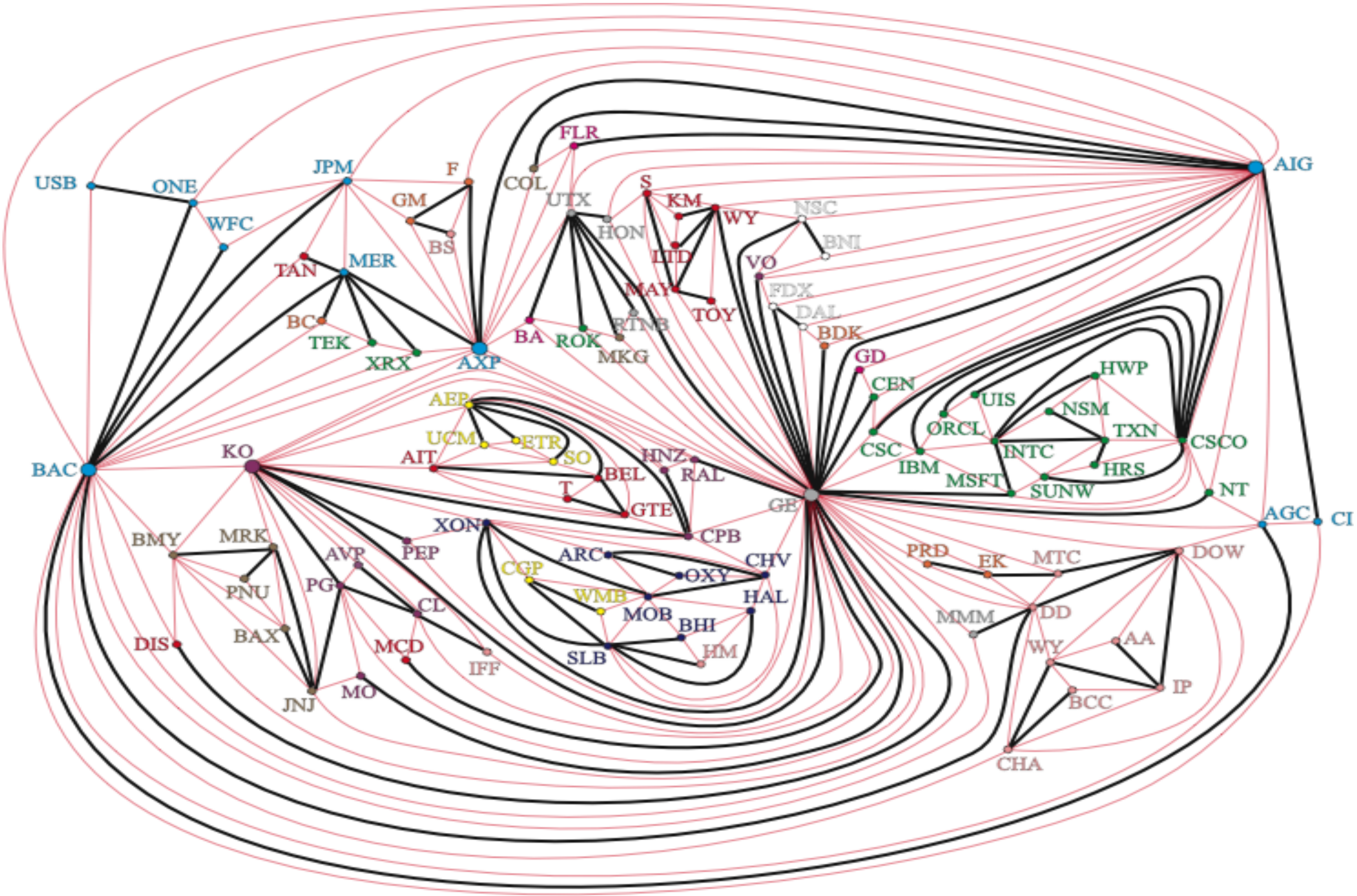
The 4-cliques present



4-cliques

AIG	BAC	AXP	MER
AIG	AXP	MER	RD
IBM	BAC	AXP	MER
IBM	TXN	AXP	MER
IBM	TXN	MOT	MER
AXP	MER	SLB	RD
MER	SLB	RD	OXY

Planar Maximally Filtered Graph in a portfolio of stocks



N=100 (NYSE)
daily returns
1995-1998
T=1011

M. Tumminello et al PNAS USA 102, 10421 (2005)

An assessment of the statistical reliability of links by using bootstrap replicas

Data Set

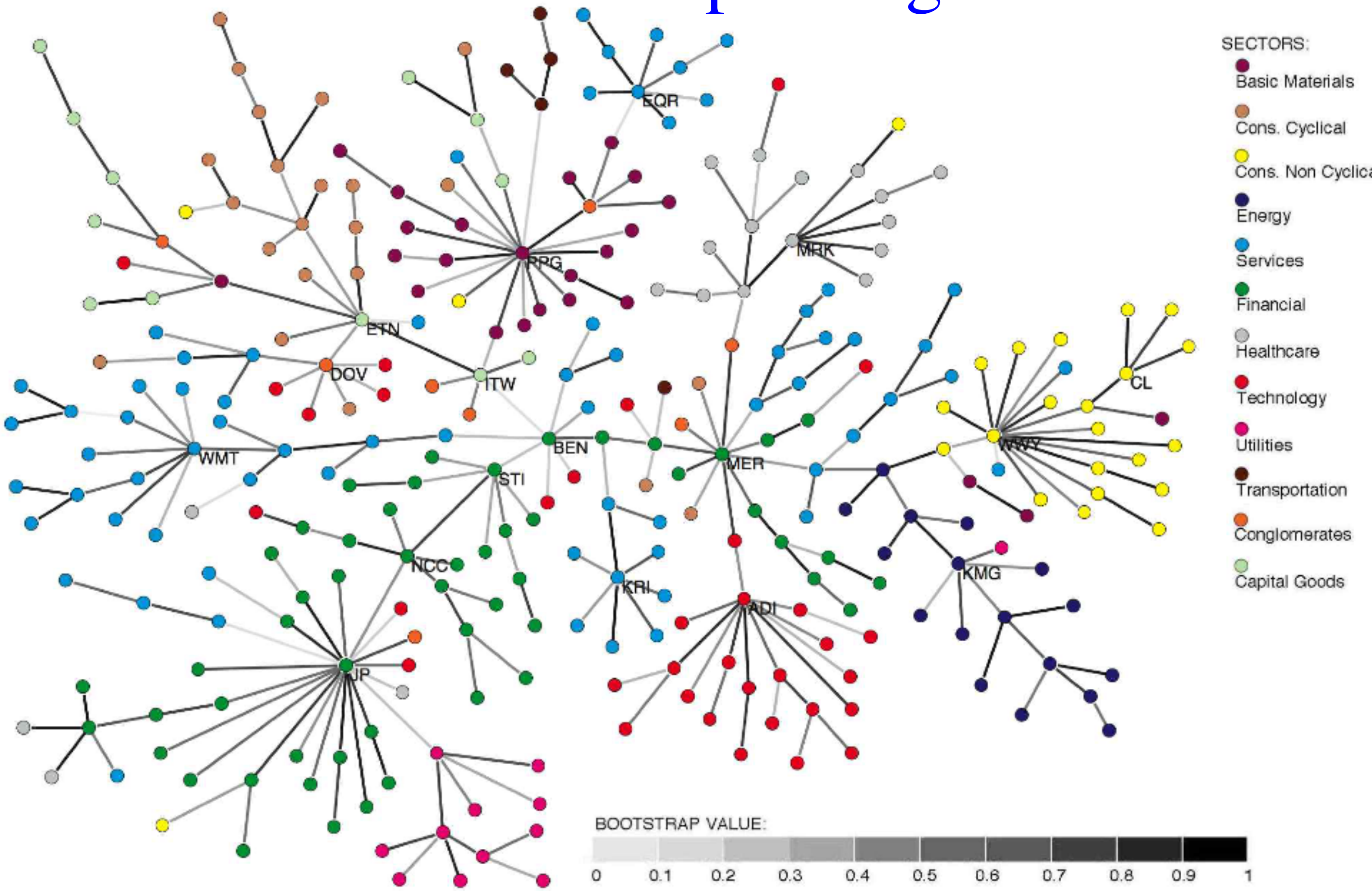
	V_1	V_2	V_3	…	V_n
t_1	0.113	1.123	-0.002	…	0.198
t_2	1.567	0.789	0.842	…	-0.234
t_3	1.065	-1.962	0.567	…	1.785
t_4	1.112	0.998	-0.424	…	2.735
t_5	-0.211	0.312	-0217	…	0.587
…	…	…	…	…	…
T	0.479	-1.828	-2.041	…	-0.193

Pseudo-replicate Data Set

V_1	V_2	V_3	…	V_n
1.567	0.789	0.842	…	-0.234
0.113	1.123	-0.002	…	0.198
1.065	-1.962	0.567	…	1.785
0.113	1.123	-0.002	…	0.198
0.479	-1.828	-2.041	…	-0.193
…	…	…	…	…
0.479	-1.828	-2.041	…	-0.193

M surrogated data matrices are constructed, e.g. M=1000. M proximity based networks are obtained and the presence of a specific link is checked in all bootstrap replicas.

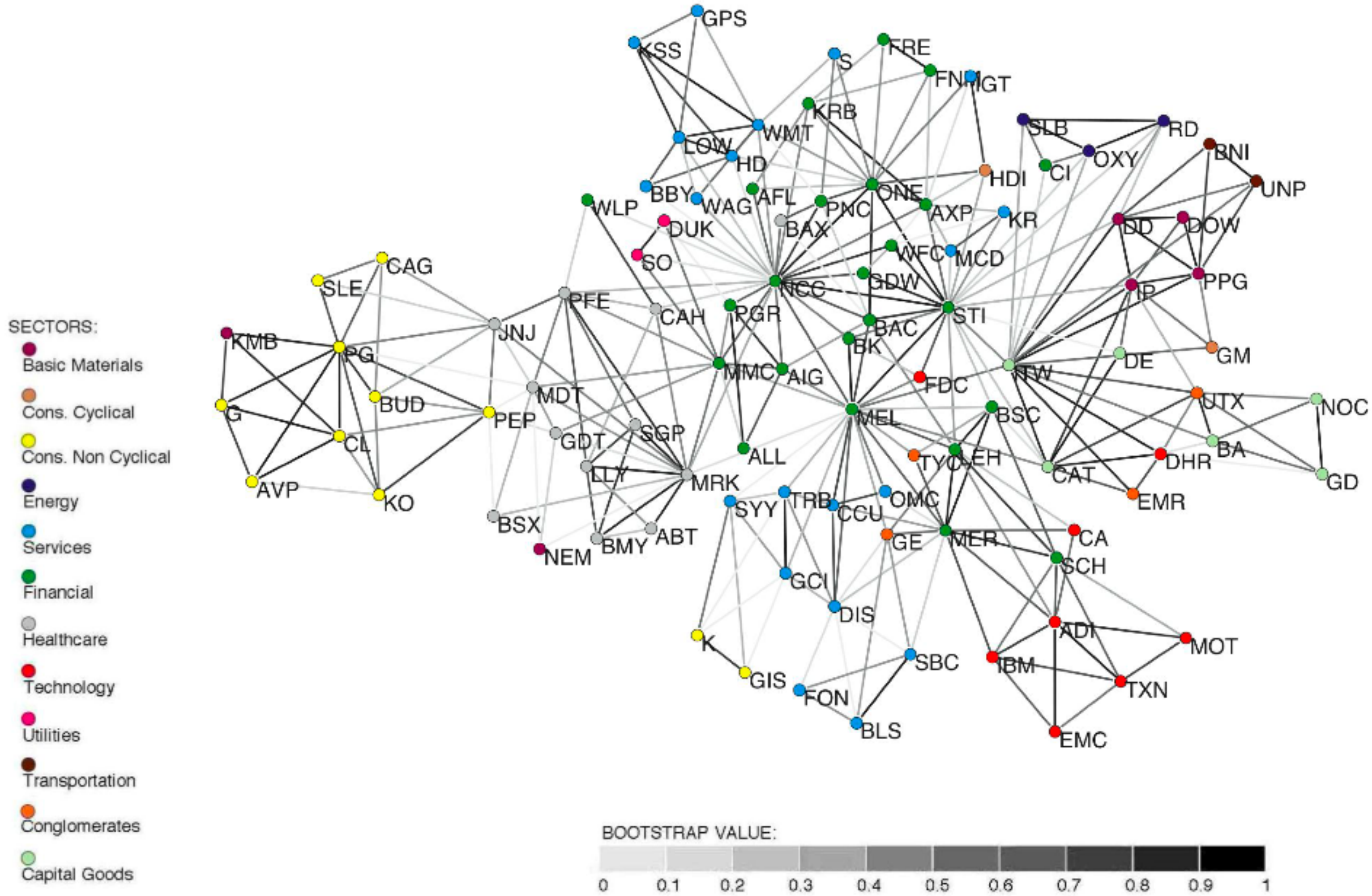
Statistical reliability of the minimum spanning tree



$N_V=300$
Stocks of US equity markets
2001-2003
$T=748$

Tumminello, Michele, et al. "Spanning trees and bootstrap reliability estimation in correlation-based networks." International Journal of Bifurcation and Chaos 17.07 (2007): 2319-2329.

Statistical reliability of the planar maximally filtered graph



NYSE

$N = 100$

daily

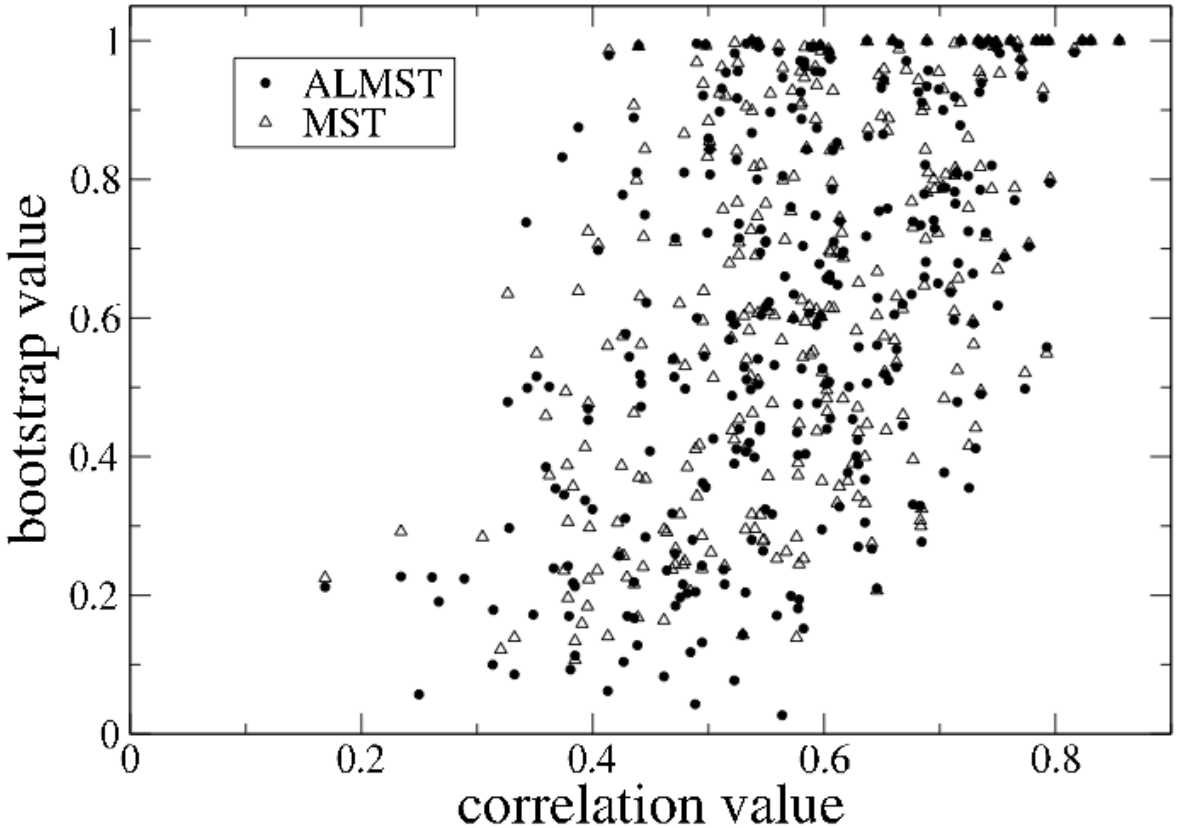
returns

2002

$T = 256$

What are the determinants of the bootstrap value of a link?

- The value of the proximity (correlation) between the two nodes;
- The topology of the proximity based network.



N_V=300
Stocks of
US equity
markets
2001-2003
T=748

Partial correlations

The partial correlation coefficient

$$\rho(X,Y:Z)$$

between variables X and Y conditioned on the variable Z is the Pearson correlation coefficient between the residuals of X and Y that are uncorrelated with Z

The partial correlation coefficient can be expressed as

$$\rho(X,Y:Z) = \frac{\rho(X,Y) - \rho(X,Z)\rho(Y,Z)}{\sqrt{\left[1-\rho^2(X,Z)\right]\left[1-\rho^2(Y,Z)\right]}}$$

A study of returns time series of stocks traded in a financial market¶

One quantity that can be investigated is

$$d(X,Y:Z) \equiv \rho(X,Y) - \rho(X,Y:Z)$$

This is an estimation of the correlation influence of Z on the correlation of pair of elements X and Y

It should be noted that *d*(X,Y:Z) assumes non negligible values only when $\rho(X,Y)$ is significantly different from zero.

¶Kenett, Dror Y., et al. "Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market." PloS one 5.12 (2010): e15032.

The number of d(X,Y:Z) elements is cubic in N.
In fact different elements are N (N-1) (N-2)/2

We therefore investigate the overall effect of stock Z on correlation of stock X with all other stocks except Z.

Specifically, we investigate

$$d(X:Z) = \langle d(X,Y:Z) \rangle_{Y \neq X,Z}$$

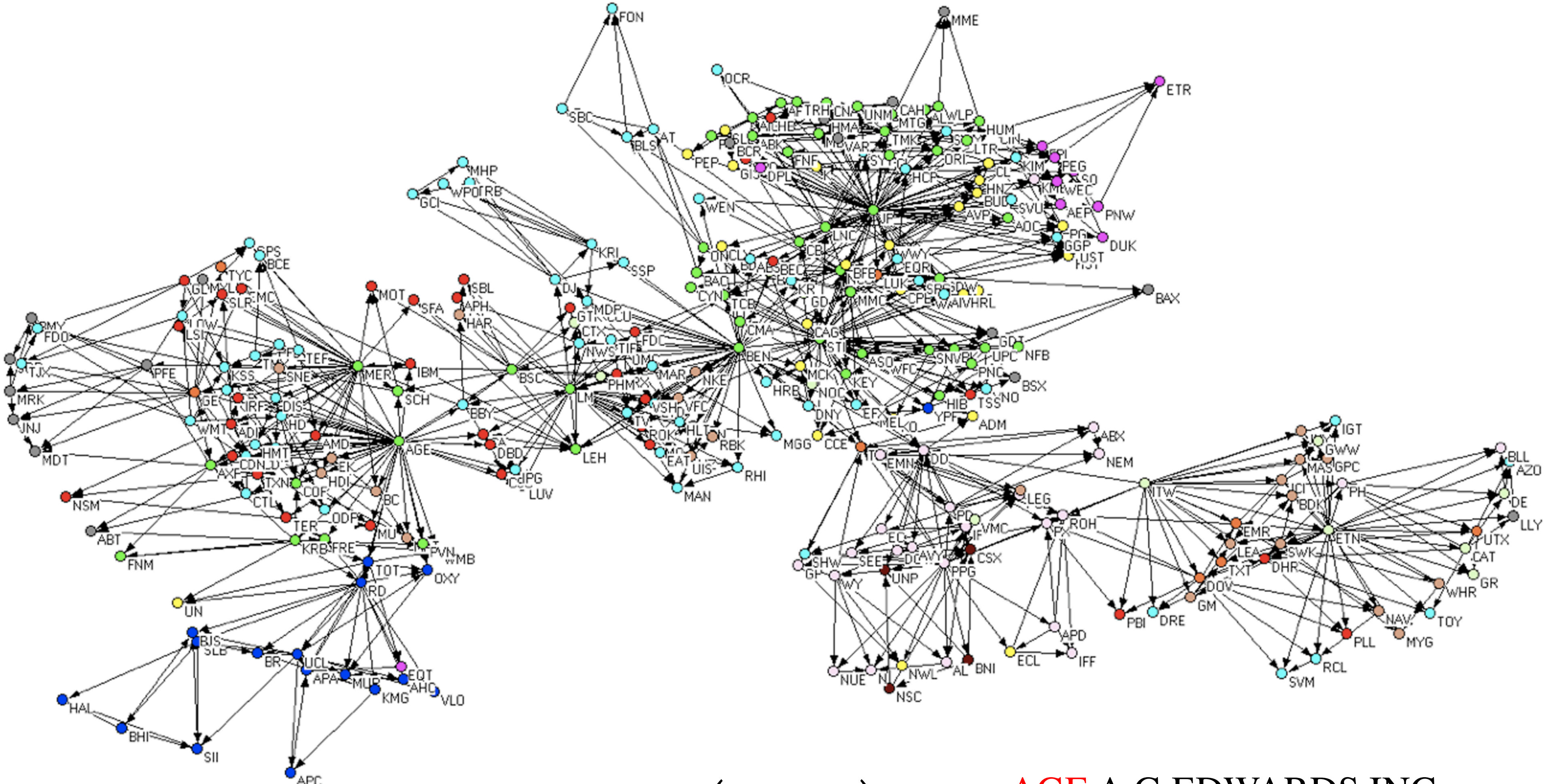
Authors use this directed similarity measure to obtain a Partial Correlation Planar Graph

The investigated set:

index	tick	sector	subsector
1	GE	Conglomerates	Conglomerates
2	PFE	Healthcare	Major Drugs
3	WMT	Services	Retail Department & Discount
4	AIG	Financial	Insurance Prop. & Casualty
5	IBM	Technology	Computer Hardware
6	KO	Consumer_Non_Cyclical	Beverages Non-Alcoholic
7	JNJ	Healthcare	Major Drugs
8	PG	Consumer_Non_Cyclical	Personal & Household Products
9	MRK	Healthcare	Major Drugs
10	BAC	Financial	Money Center Banks
11	WFC	Financial	Money Center Banks
12	SBC	Services	Communication Services
13	FNM	Financial	Consumer Financial Services
14	HD	Services	Retail Home Improvement
15	PEP	Consumer_Non_Cyclical	Beverages Non-Alcoholic
16	LLY	Healthcare	Major Drugs
17	BUD	Consumer_Non_Cyclical	Beverages Alcoholic
18	ABT	Healthcare	Major Drugs
19	BMY	Healthcare	Major Drugs
20	AXP	Financial	Consumer Financial Services
21	MER	Financial	Investment Services
22	MDT	Healthcare	Medical Equipment & Supplies
23	UTX	Conglomerates	Conglomerates
24	BLS	Services	Communication Services
25	ONE	Financial	Regional Banks
………………			
291	AIV	Services	Real Estate Operations
292	VSH	Technology	Electronics Instruments & Controls
293	BEC	Technology	Scientific & Technical Instr.
294	BC	Consumer_Cyclical	Recreational Products
295	MYG	Consumer_Cyclical	Appliance & Tool
296	HCP	Services	Real Estate Operations
297	EQT	Utilities	Natural Gas Utilities
298	IRF	Technology	Semiconductors
299	CYN	Financial	Regional Banks
300	AL	Basic_Materials	Metal Mining

300 highly capitalized stocks traded in the US equity markets during 2001-2003

Partial correlation planar graph at the level of single stock



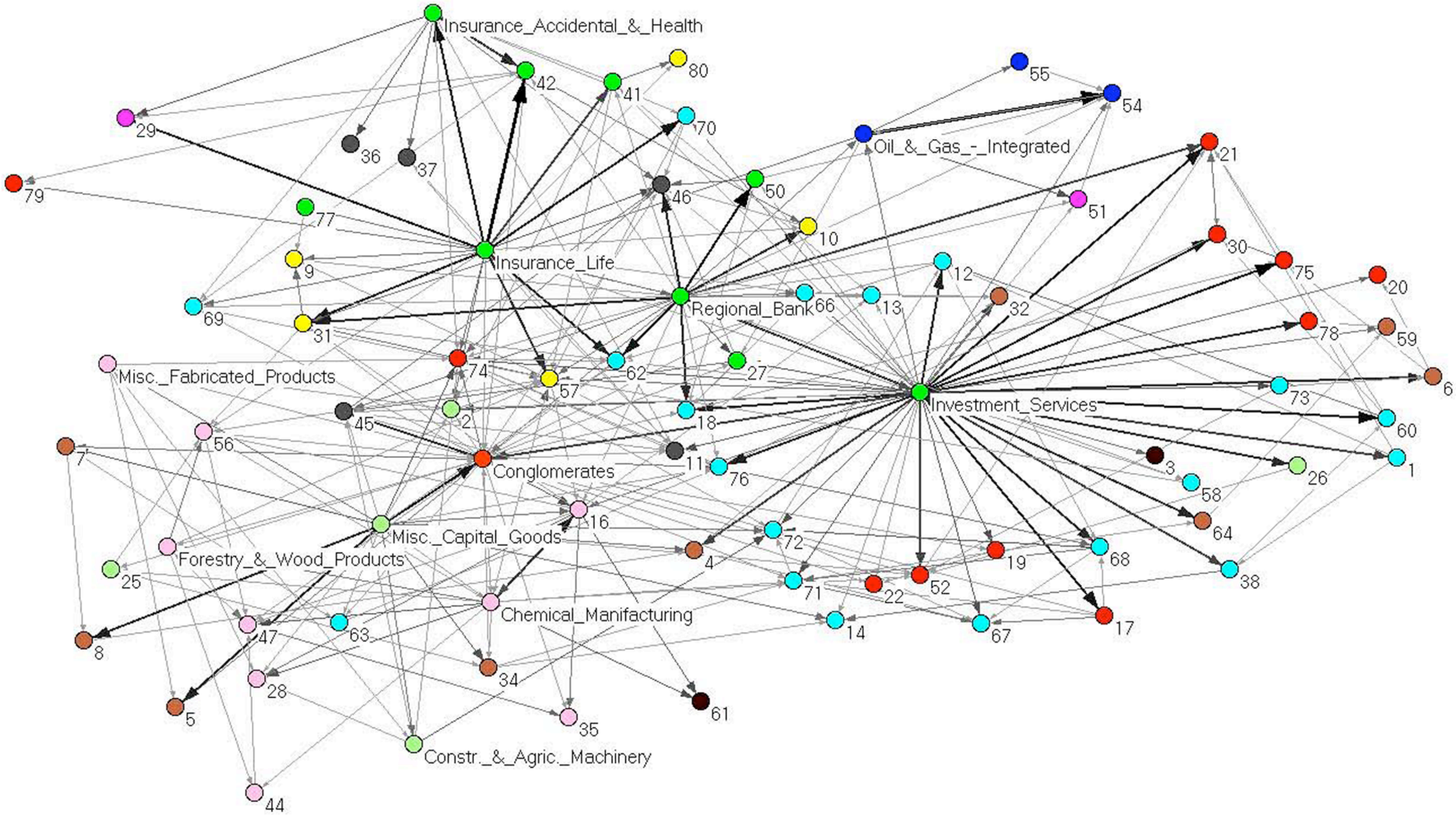
$$d(X:Z) \neq d(Z:X)$$

AGE A.G.EDWARDS INC

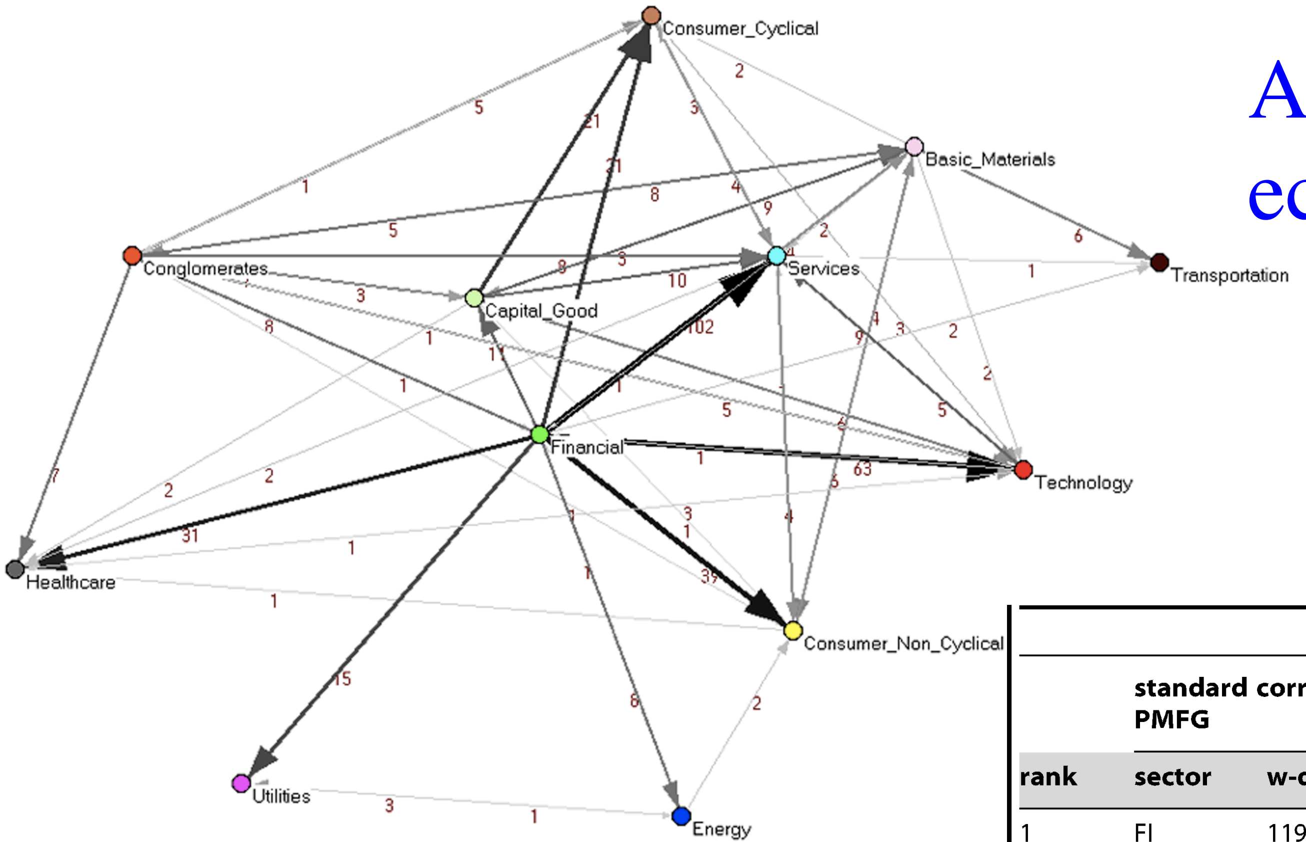
BEN FRANKLIN RESOURCES

JP JEFFERSON-PILOT CORP

The Partial Correlation Planar Graph at the level of economic subsectors



At the level of economic sectors



	standard correlation: PMFG		partial correlation: PCPG		
rank	sector	w-degree	sector	w-outdegree	w-indegree
1	FI	119	FI	304	4
2	SE	85	CG	56	17
3	BM	60	CO	38	22
4	CO	55	BM	26	25
5	CG	53	SE	16	136
6	TE	51	TE	12	87
7	CC	49	CC	8	52
8	CN	29	EN	6	9
9	HE	24	CN	6	53
10	EN	15	HE	3	44
11	UT	11	UT	1	18
12	TR	9	TR	0	9

doi:10.1371/journal.pone.0015032.t003

Granger causality networks[¶]

The authors perform a linear Granger causality test. Time series *j* is said to "Granger-cause" time series *i* if past values of *j* contain information able to predict *i* above and beyond information contained in past values of *i* alone.

[¶]Billio, Monica, Mila Getmansky, Andrew W. Lo, and Loriana Pelizzon. "Econometric measures of connectedness and systemic risk in the finance and insurance sectors." Journal of Financial Economics 104, no. 3 (2012): 535-559.

They considered the linear model

$$R^i_{t+1} = a^i R^i_t + b^{ij} R^j_t + e^i_{t+1}$$
$$R^j_{t+1} = a^j R^j_t + b^{ji} R^i_t + e^j_{t+1}$$

Authors conclude that j Granger-causes i when b^{ij} is (statistically) different from zero and i Granger-causes j when b^{ji} is (statistically) different from zero.

Authors investigated monthly returns data for hedge funds, broker/dealers, banks and insurers. The investigated set comprises 100 top financial institutions (25 for each category) during the period from January 1994 through December 2008.

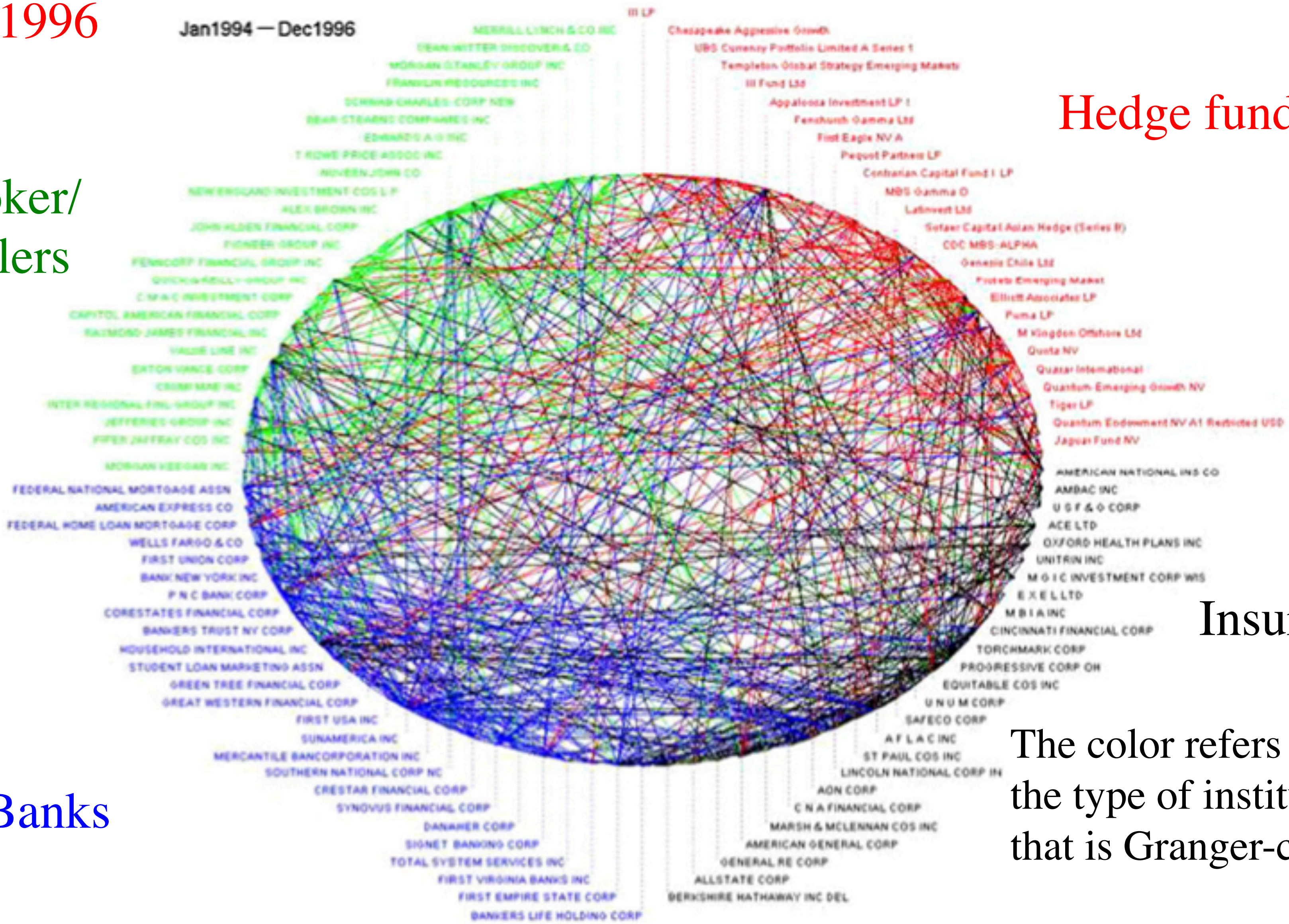
Heteroskedasticity is taken into account by using a GARCH(1,1) baseline model of returns.

Jan 1994-
Dec 1996

M. Billio et al. / Journal of Financial Economics 104 (2012) 535–559

Hedge funds

Broker/
dealers



Insurers

Banks

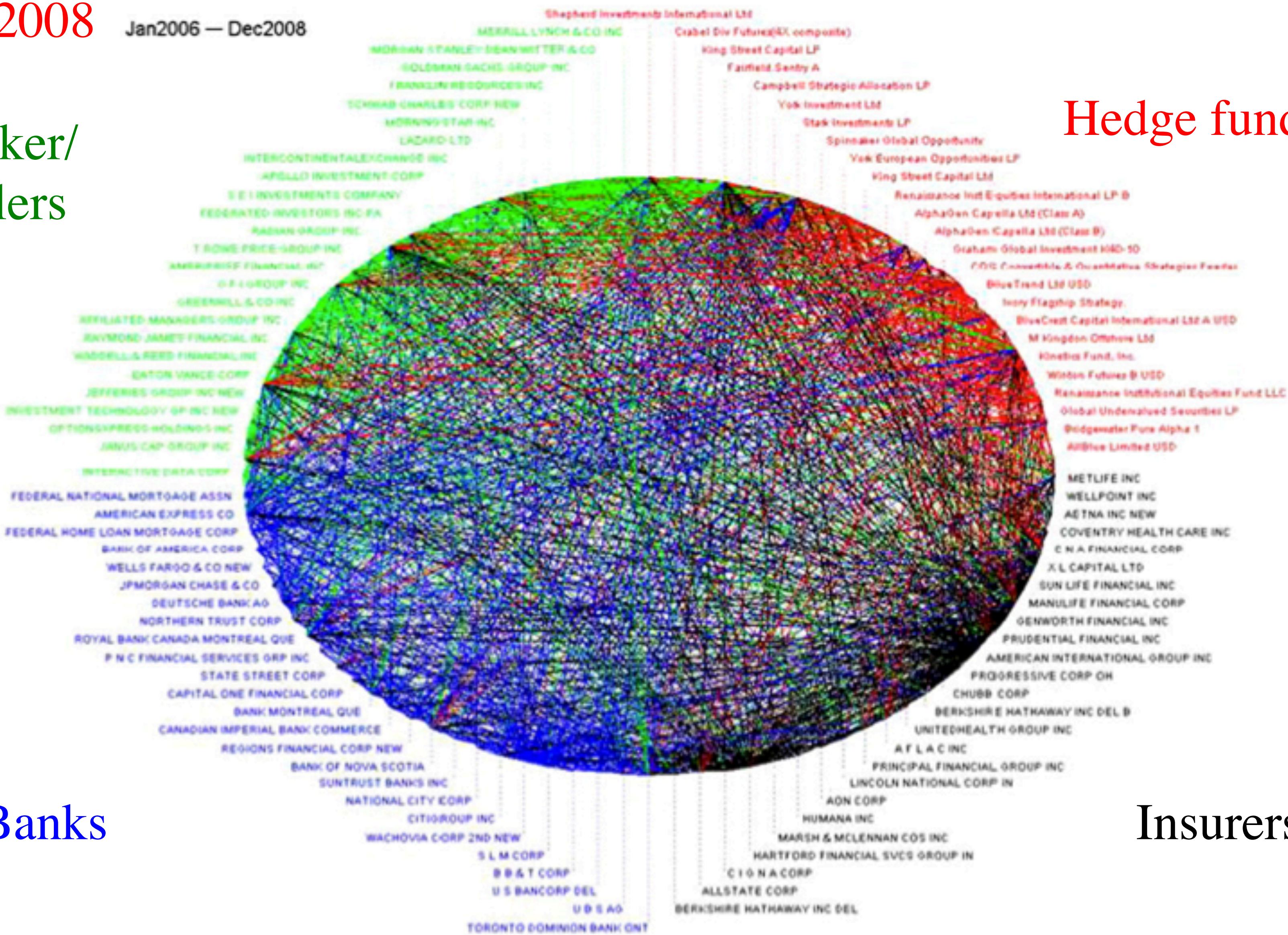
The color refers to the type of institution that is Granger-causing

Fig. 2. Network diagram of linear Granger-causality relationships that are statistically significant at the 5% level among the monthly returns of the 25 largest (in terms of average market cap and AUM) banks, broker/dealers, insurers, and hedge funds over January 1994 to December 1996. The type of institution causing the relationship is indicated by color: green for broker/dealers, red for hedge funds, black for insurers, and blue for banks. Granger-causality relationships are estimated including autoregressive terms and filtering out heteroskedasticity with a GARCH(1,1) model.

Jan 2006-
Dec 2008

Broker/
dealers

Hedge funds



Banks

Insurers

Fig. 3. Network diagram of linear Granger-causality relationships that are statistically significant at the 5% level among the monthly returns of the 25 largest (in terms of average market cap and AUM) banks, broker/dealers, insurers, and hedge funds over January 2006 to December 2008. The type of institution causing the relationship is indicated by color: green for broker/dealers, red for hedge funds, black for insurers, and blue for banks. Granger-causality relationships are estimated including autoregressive terms and filtering out heteroskedasticity with a GARCH(1,1) model.

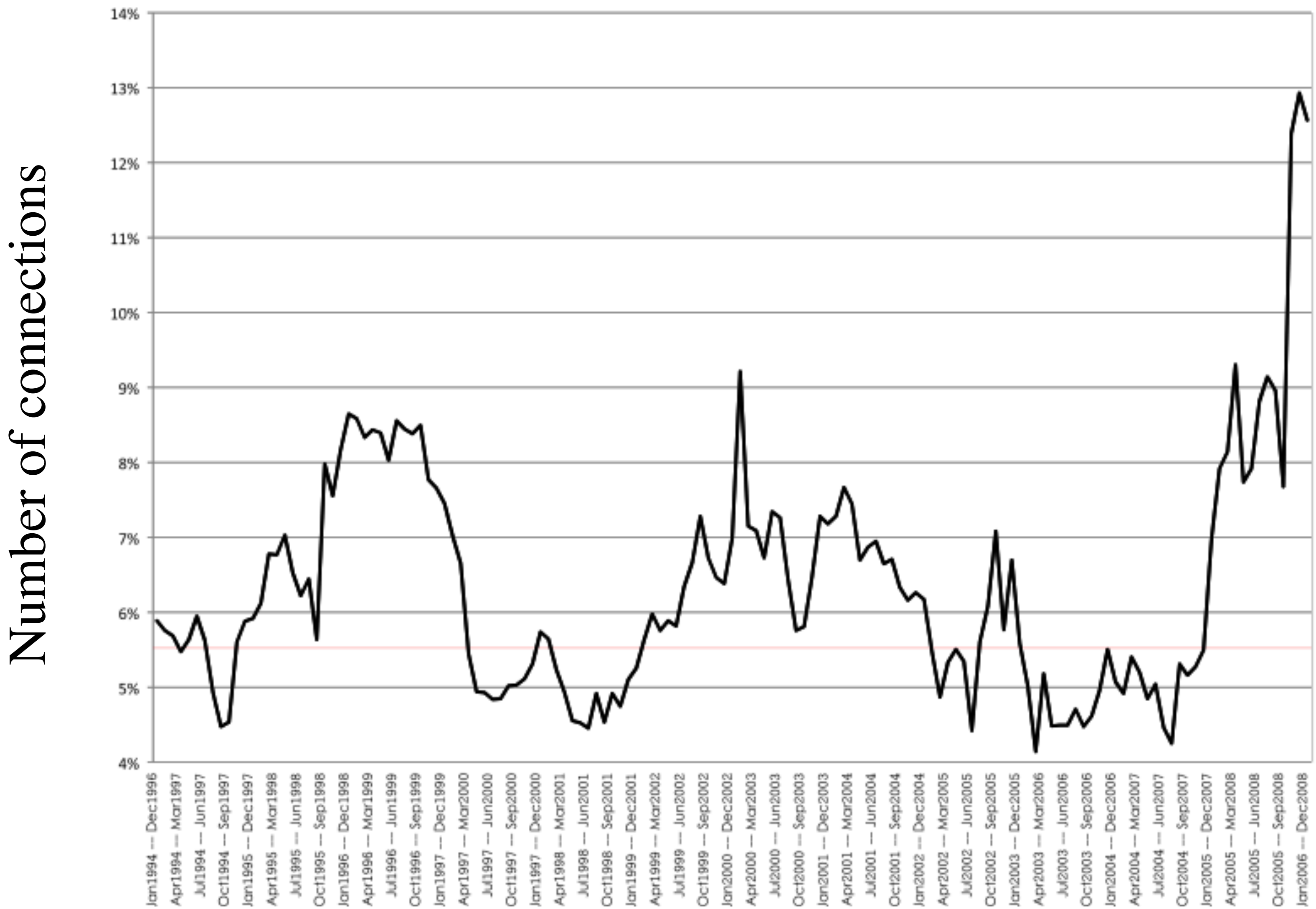


Fig. 4. Number of connections as a percentage of all possible connections. The time series of linear Granger-causality relationships (at the 5% level of statistical significance) among the monthly returns of the largest 25 banks, broker/dealers, insurers, and hedge funds (as determined by average AUM for hedge funds and average market capitalization for broker/dealers, insurers, and banks during the time period considered) for 36-month rolling-window sample periods from January 1994 to December 2008. The number of connections as a percentage of all possible connections (our DGC measure) is depicted in black against 0.055, the 95% of the simulated distribution obtained under the hypothesis of no causal relationships depicted in a thin line. The number of connections is estimated for each sample including autoregressive terms and filtering out heteroskedasticity with a GARCH(1,1) model.

Co-integration networks

Physica A 400 (2014) 168–185

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Cointegration analysis and influence rank—A network approach to global stock markets

Chunxia Yang*, Yanhua Chen, Lei Niu, Qian Li

Nanjing University of Information Science and Technology, Nanjing 210044, China

Physica A 402 (2014) 245–254

ntents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Cointegration-based financial networks study in Chinese stock market

Chengyi Tu

School of Automation Engineering, University of Electronic Science and Technology of China, 611731 Chengdu, Sichuan, PR China

Edge filtering is also relevant in networks

Several networks are pretty dense and it is quite difficult to detect their internal structures.

One recent approach[¶] able to detect internal structures of networks is the approach of statistically validated networks.

In statistically validated networks the scientific question is:

Is it possible to detect interaction among nodes of the network that are over- expressed or under-expressed with respect to a null hypothesis taking into account the heterogeneity of the system?

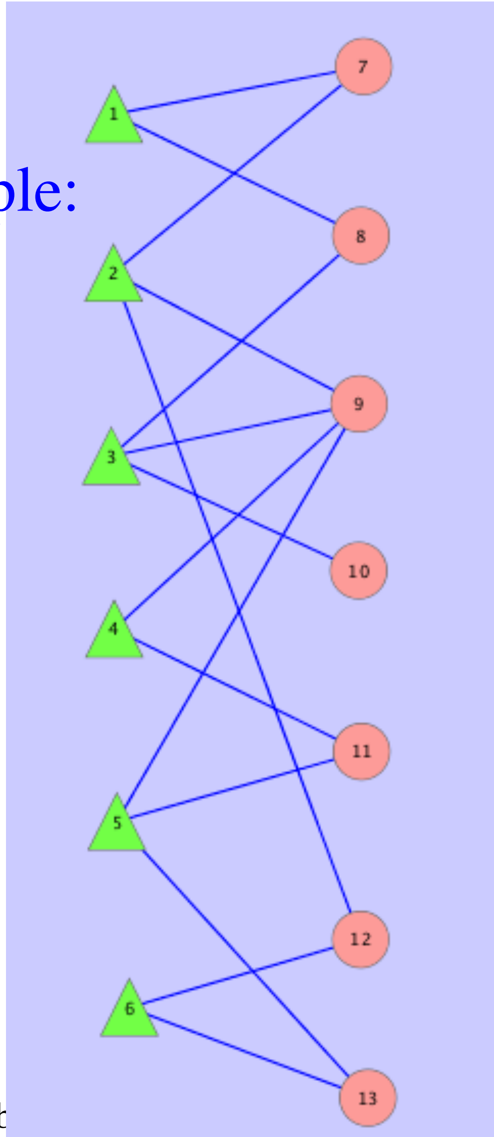
[¶] Tumminello M, Miccichè S, Lillo F, Piilo J, Mantegna RN (2011) Statistically Validated Networks in Bipartite Complex Systems. PLoS ONE 6(3): e17994. doi:10.1371/journal.pone.0017994

In several cases the problem of statistically validating a link can be mapped into a urn problem

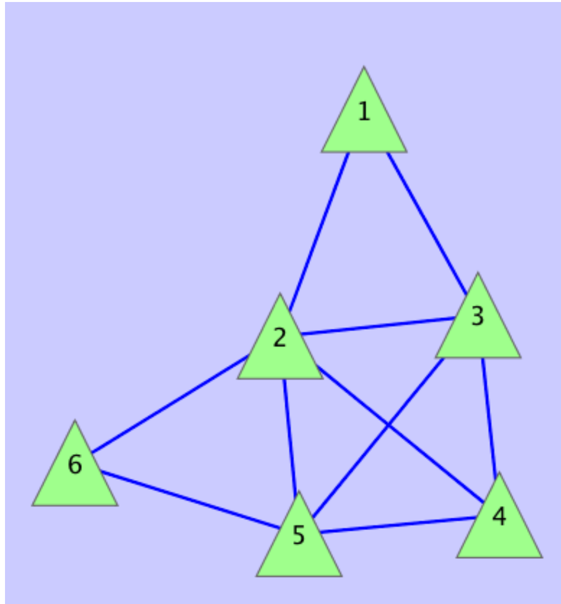
Lending banks

Packages

Example:



Projected network of lending banks



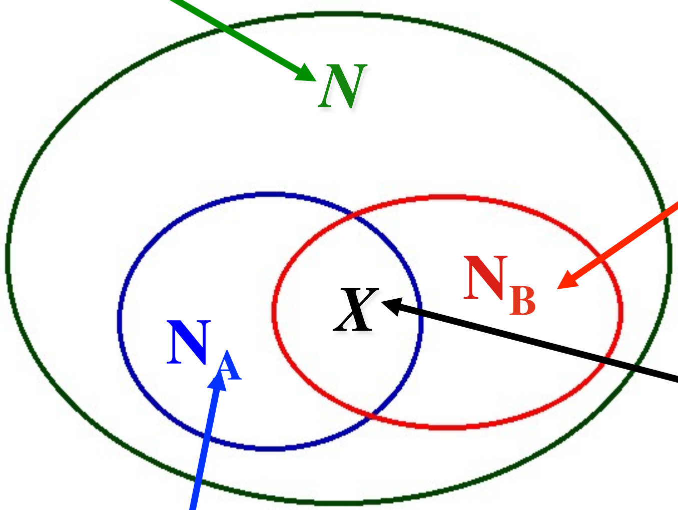
The investigated system concerns syndicated loans.

The database is the DealScan database of Thomson Reuters

A statistical validation of co-occurrence

Suppose there are **N** loan packages in the investigated set. Suppose we are interested to evaluate against a null hypothesis the co-occurrence of lending banks in the same package. Let us call $\mathbf{N_A}$ the number of packages that bank **A** has subscribed and $\mathbf{N_B}$ the number of packages that bank **B** has subscribed. Let us call X the co-occurrence of the presence of both banks in loan packages.

Total # of packages



# of packages N_B subscribed by bank B

# of co-occurrence of banks A and B in subscribed packages

# of packages N_A subscribed by bank A

The question is:
what is the probability of X under the null hypothesis of random matching?

The probability that banks **A** and **B** are both subscribing X packages is given by the hypergeometric distribution

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

Expected number of co-occurrence: $$\langle X \rangle = \sum x \; P(x \mid N, N_A, N_B)$$

It is therefore possible to associate a p-value to an empirically observed value (indeed it is possible to perform a two tails test)

p-value associated with a detection of co-occurrence $\geq$ X:

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

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

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

Corrections for multiple hypotheses testing, and network construction

We can therefore statistically validate a link between two vertices (in the present case two banks) if the associated *p*-value is below a given threshold showing that the co-occurrence cannot be explained by the heterogeneity of the system taken as a null hypothesis.

By doing a two tail analysis we can also detect under-occurrence so that detecting the avoidance or minimization of interaction.

To perform the statistical validation of all pairs of vertices a large number of tests need to be performed. One therefore needs a multiple hypothesis test correction.

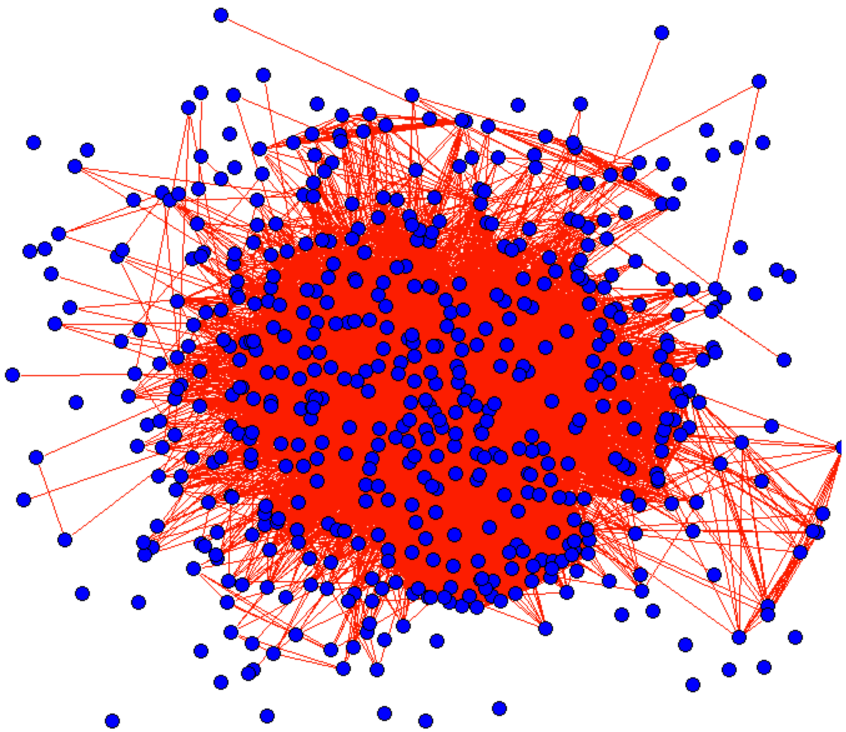
The most restrictive correction is the Bonferroni correction redefining the statistical threshold as θ=0.01/T where T is the number of tests to be done.

Another type of correction (less restrictive) is the so-called False Discovery Rate correction.

DealScan network of banks performing syndicated loans[¶]

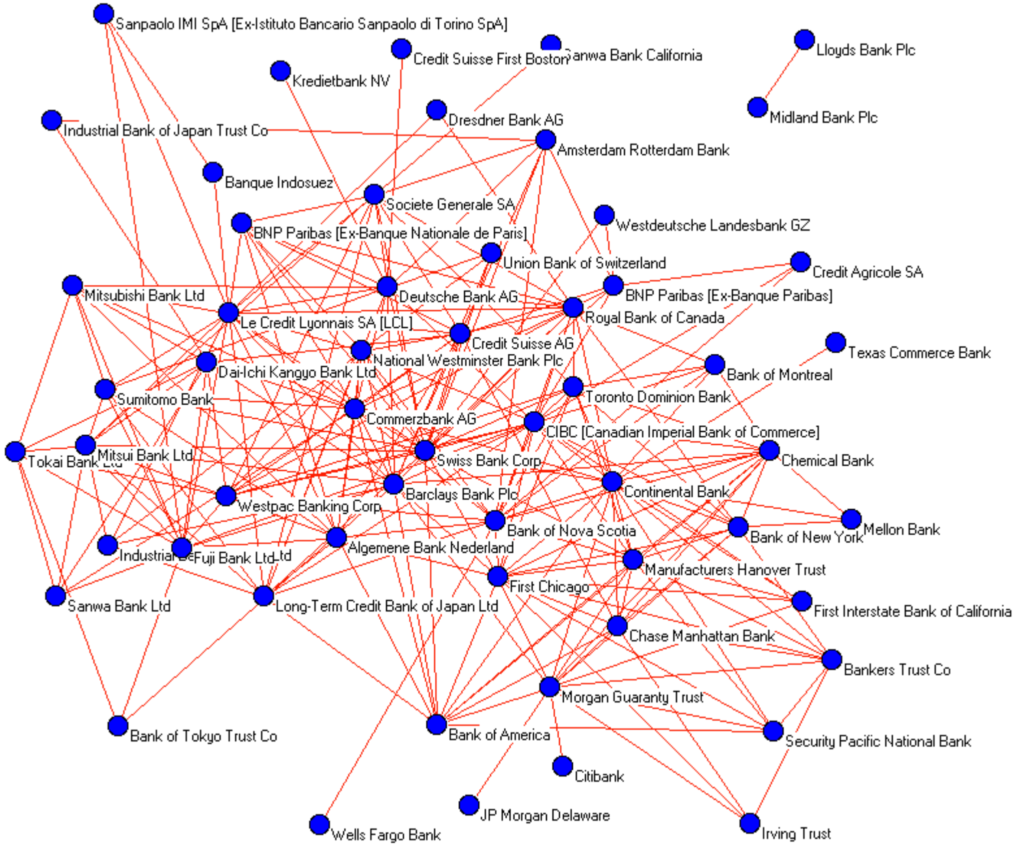
1987

Network



551 banks
13544 links

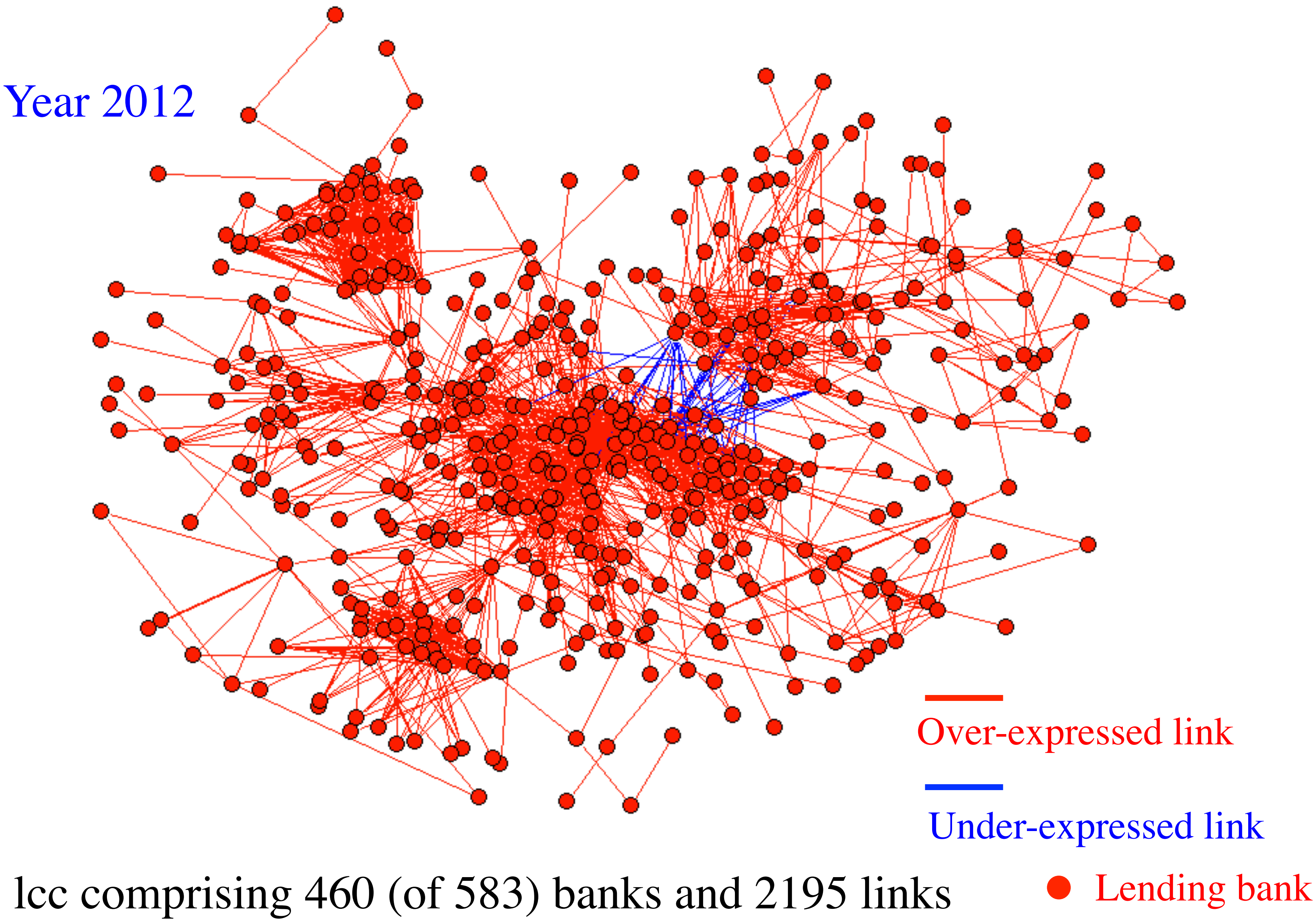
Statistically validated network



57 banks 249 links

[¶]L. Marotta, S. Miccichè and R. N. Mantegna, The evolution of the network of banks performing syndicated loans, manuscript in preparation

Statistically validated network of DealScan lending banks

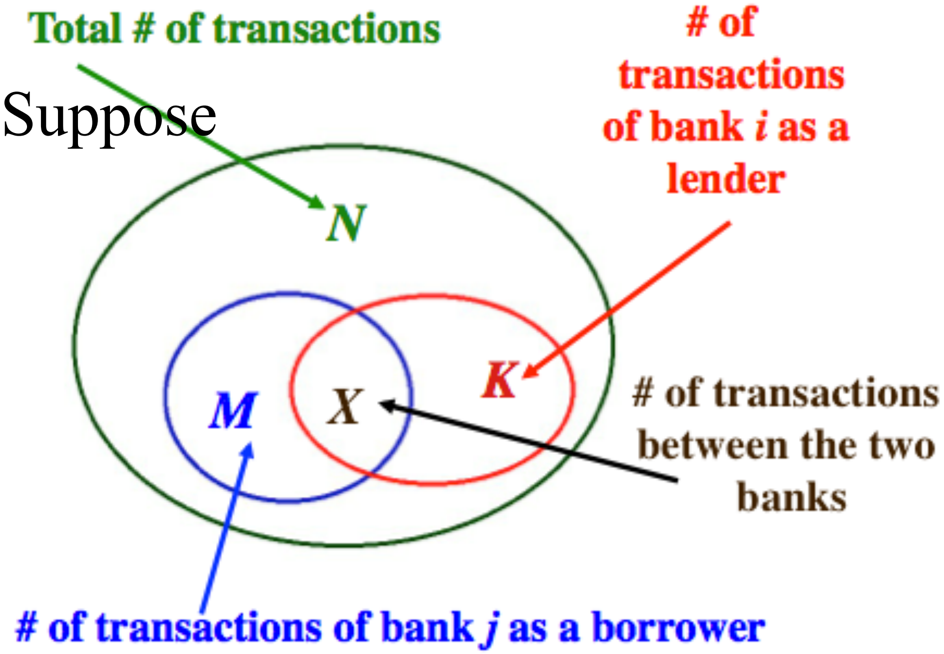


lcc comprising 460 (of 583) banks and 2195 links

The methodology of statistically validated networks is quite flexible and can be easily applied also to directed networks when the underlying network register directional events.

Example: credit relationships in the interbank market.

Suppose there are **N** credit relationships in the investigated set. Suppose We are interested to evaluate the null hypothesis of the co-occurrence of random pairing of lending and borrowing between a pair of banks. Let us call **K** the number of credits relationships of bank ***i*** as a lender and **M** the number of credit relationships of bank ***j*** as a borrower. X is the number of credit relationships with ***i*** lender and ***j*** borrower.



p-value

OVER-expression:

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

UNDER-expression:

$$p = \sum_{i=0}^{X} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{i}}$$

By using this approach we[¶] have shown that the e-MID market presents statistically validated links

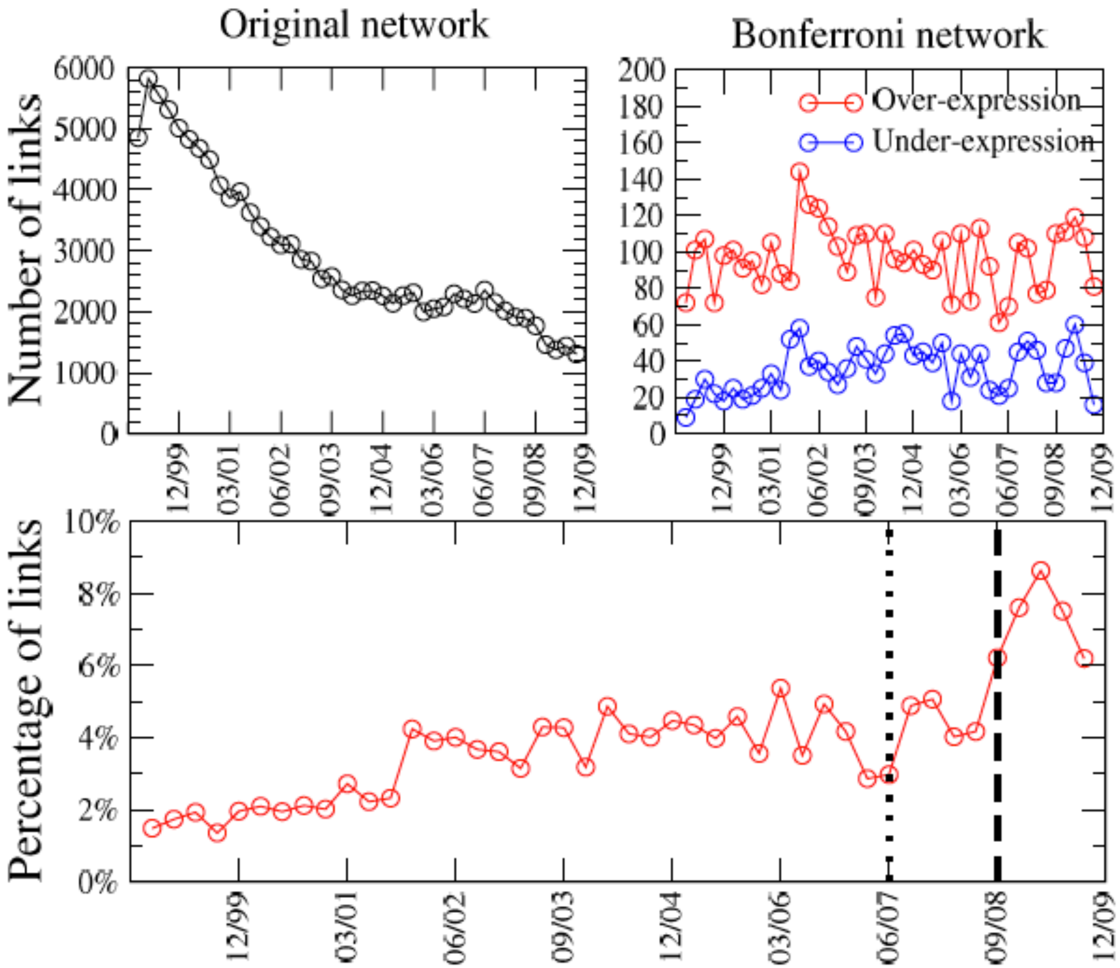


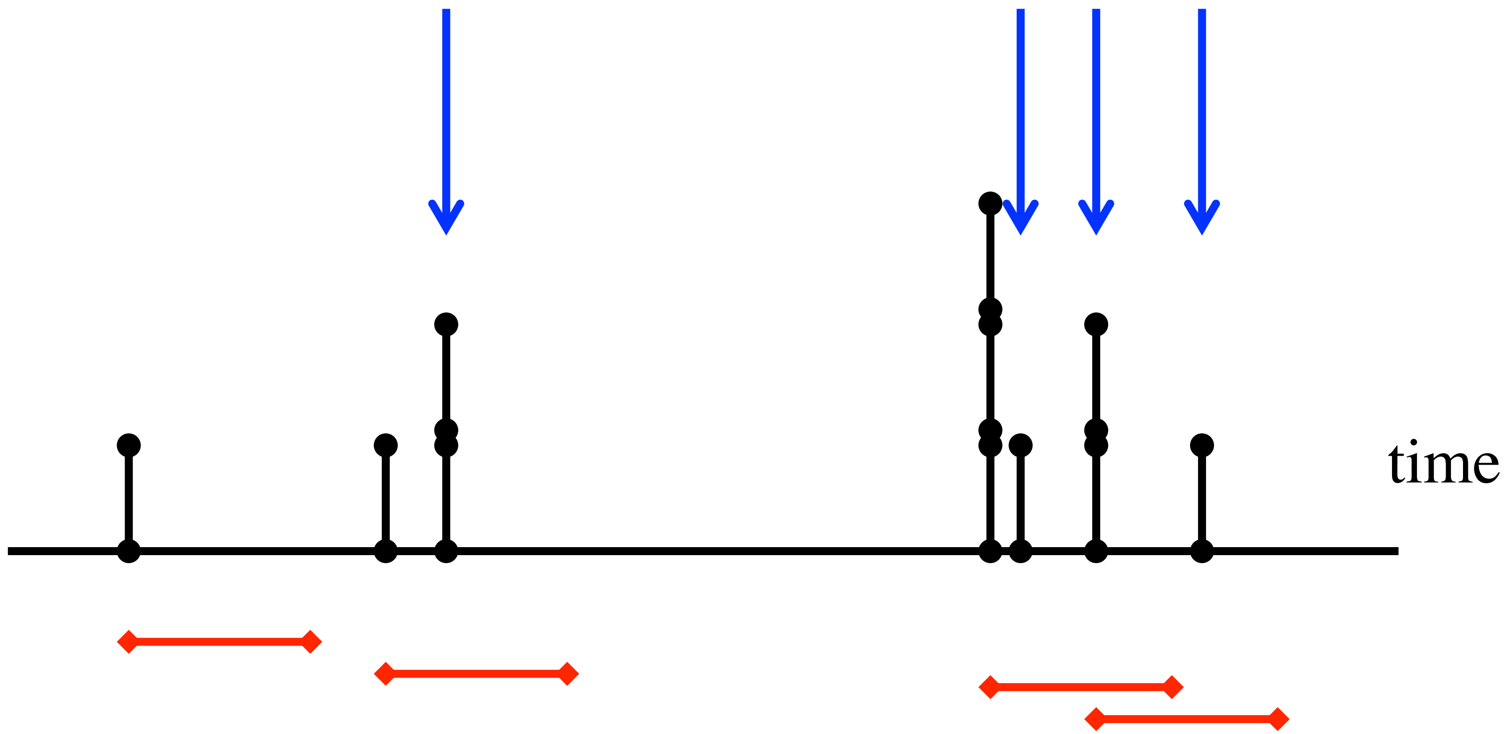
Figure 8. In the top-left panel, we show the number of links observed in the original network. In the top-right panel, the number of over-expressed links (red) and under-expressed links (blue) observed in the Bonferroni network is reported. In the bottom panel, we show the ratio between the number of over-expressed links observed in the Bonferroni and in the original network. The dotted line refers to the August 2007 market freezing, while the dashed line refers to the Lehman's bankruptcy. These data refer to the lender-aggressor dataset. The analysis is performed on the Italian segment of the e-MID market.

[¶]Hatzopoulos, V. , Iori, G., Mantegna, R. N., Miccichè, S. and Tumminello, M., Quantifying Preferential Trading in the e-MID Interbank Market (October 28, 2013). Available at SSRN: http://ssrn.com/abstract=2343647

Patterns of Algorithmic High-Frequency Trading Networks at NASDAQ OMX Helsinki

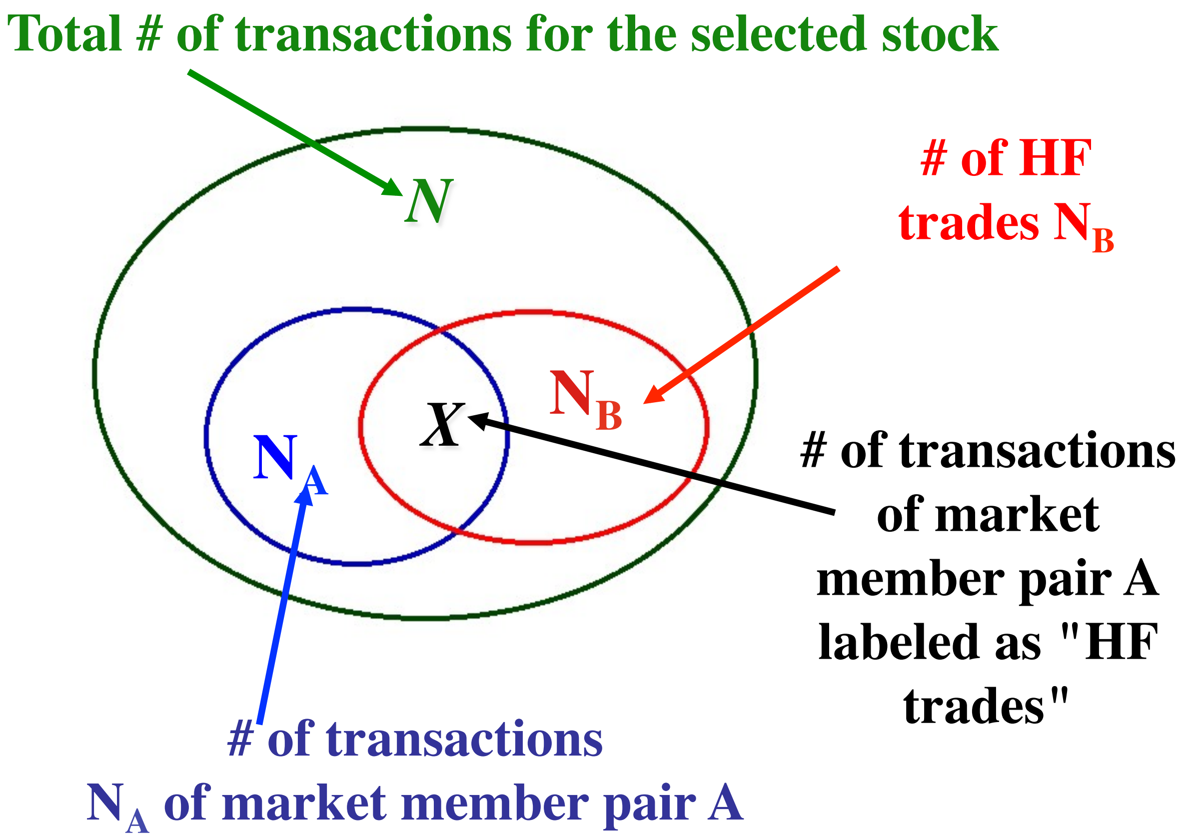
¶L. Marotta, J. Piilo and R. N. Mantegna, Patterns of Algorithmic High-Frequency Trading Networks at NASDAQ OMX Helsinki, manuscript in preparation

We label as High Frequency (HF) trades those trades occurring after a "triggering trade" within a time window τ of 1 msec or 50 msec.



A statistical validation of algorithmic HF trading pairs

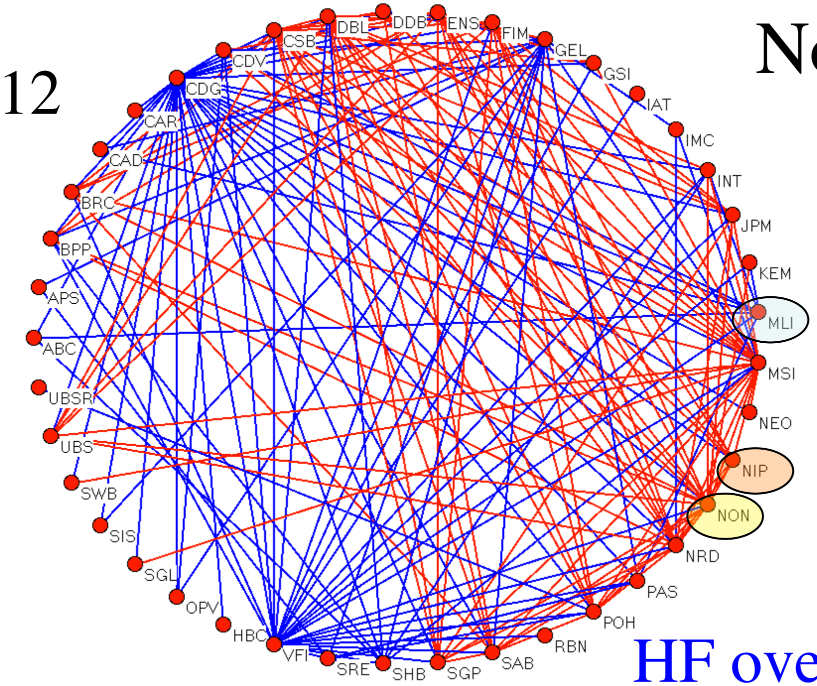
Suppose there are **N** transactions in the investigated set. Suppose we are interested to evaluate the over-expression or under-expression of "HF trades" for each pair of market members against a null hypothesis. For a given stock, let us call $\mathbf{N_A}$ the number of transactions that market member pair **A** has done and $\mathbf{N_B}$ the number of "HF trades".



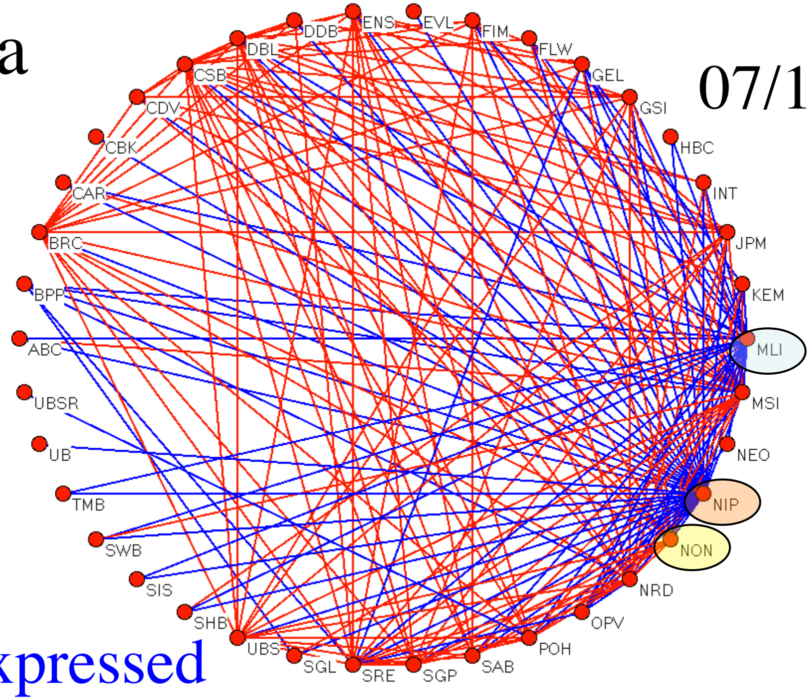
Month	Year	All ISIN	SV ISIN	SV MM	O-E ISIN	O-E MM	U-E ISIN	U-E MM	O-E pair	U-E pair
1	2012	512	47	52	47	52	26	37	1127	1110
3	2012	433	51	53	51	53	31	36	1102	1518
6	2012	438	49	51	49	50	29	38	833	1154
9	2012	471	45	47	45	45	26	35	716	988
12	2012	429	48	49	48	47	23	31	515	396
3	2013	432	49	50	49	50	28	34	650	819
6	2013	459	39	50	39	47	25	37	596	708
9	2013	550	44	49	44	42	26	42	600	823
12	2013	510	41	49	41	47	27	35	619	703

Nokia

01/12



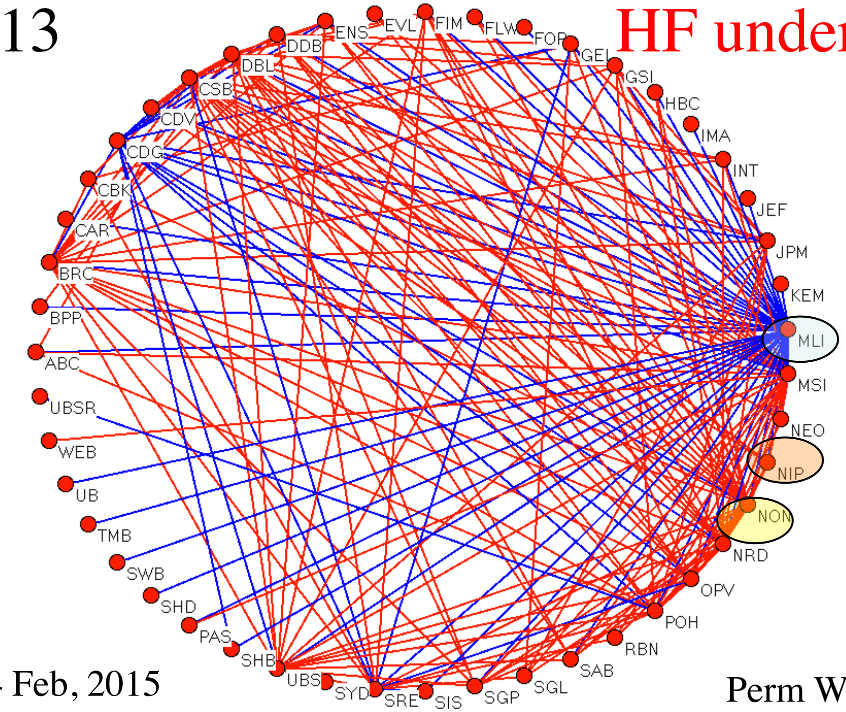
07/12



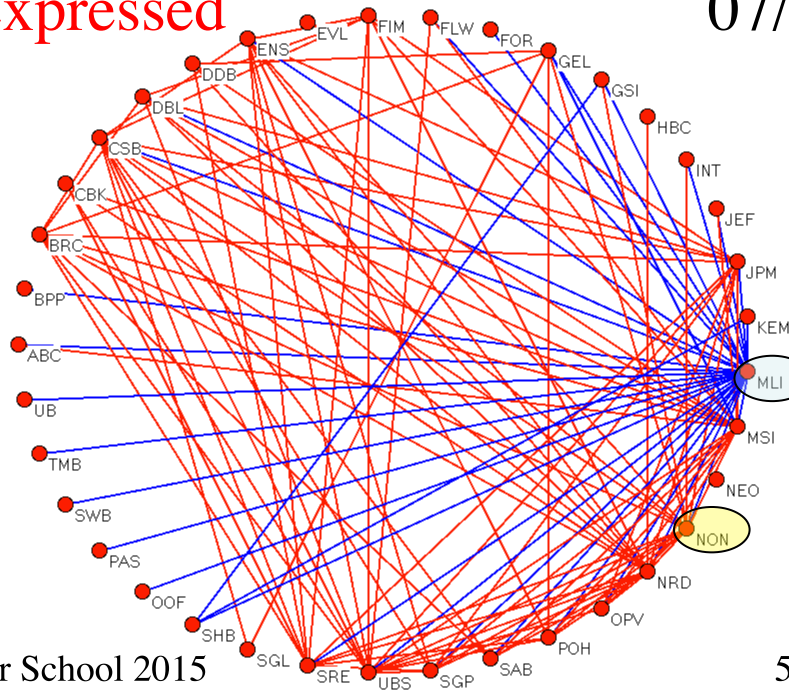
HF over-expressed

HF under-expressed

01/13



07/13

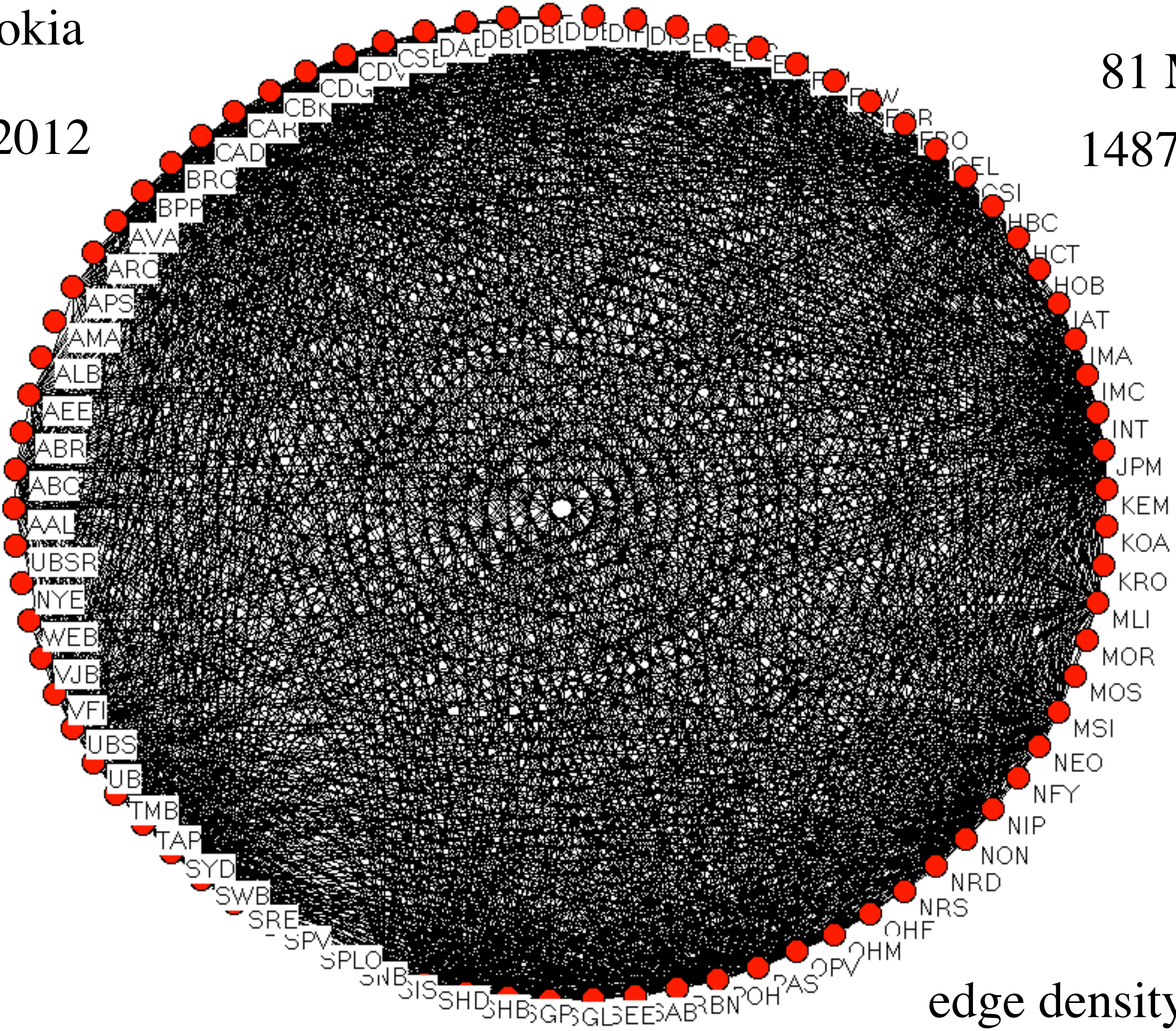


Nokia

01/2012

81 MMs

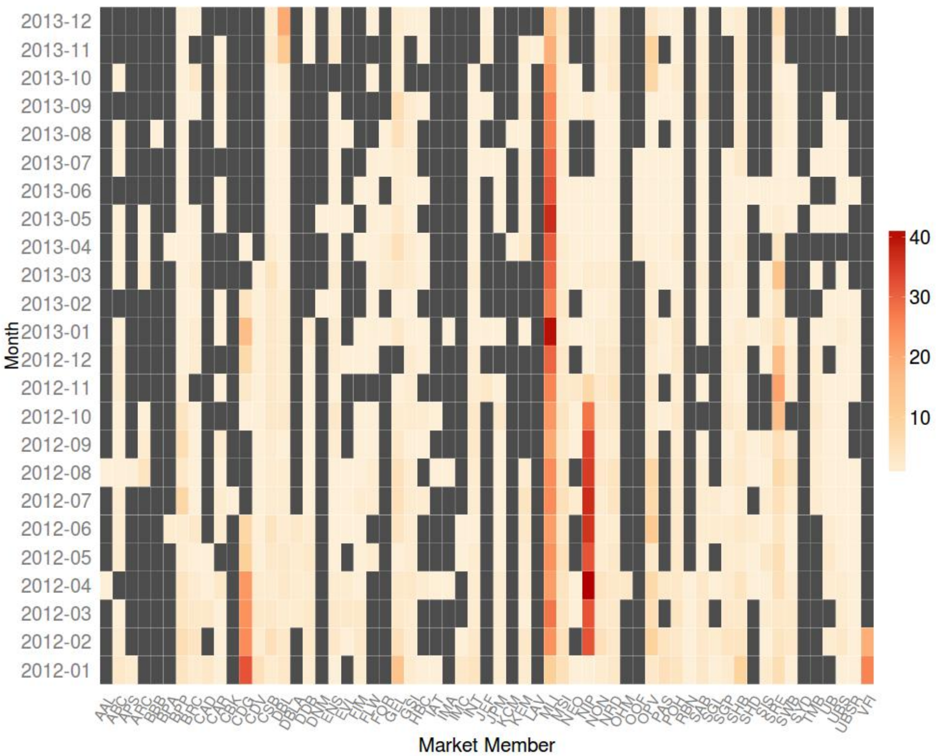
1487 edges



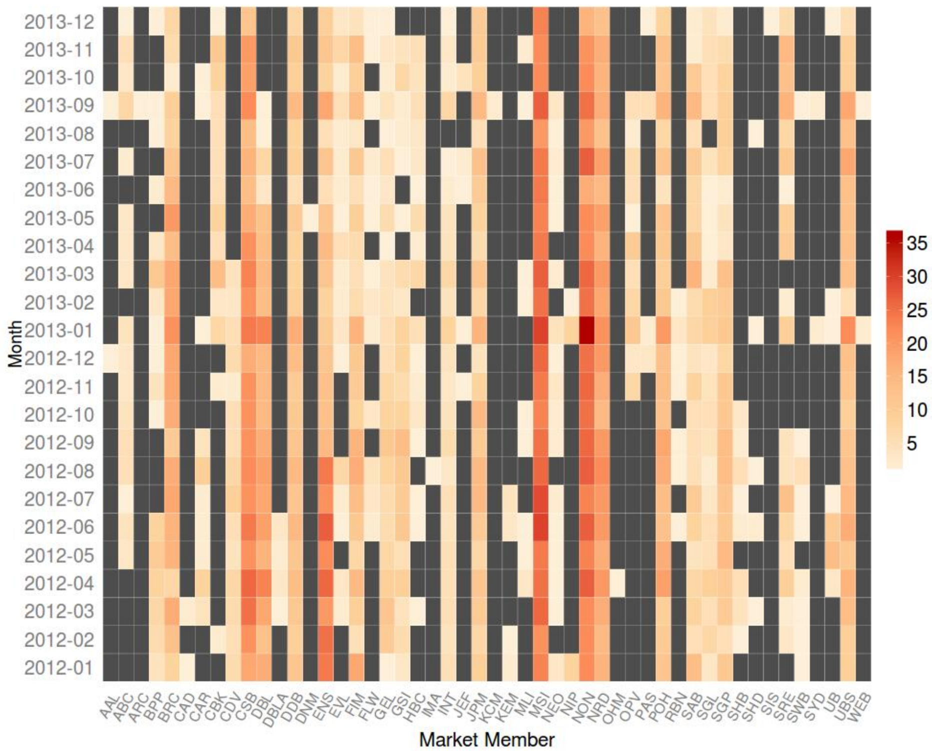
edge density=46%

MMs trading Nokia stock

Degree of MMs
(over-expressed edges)



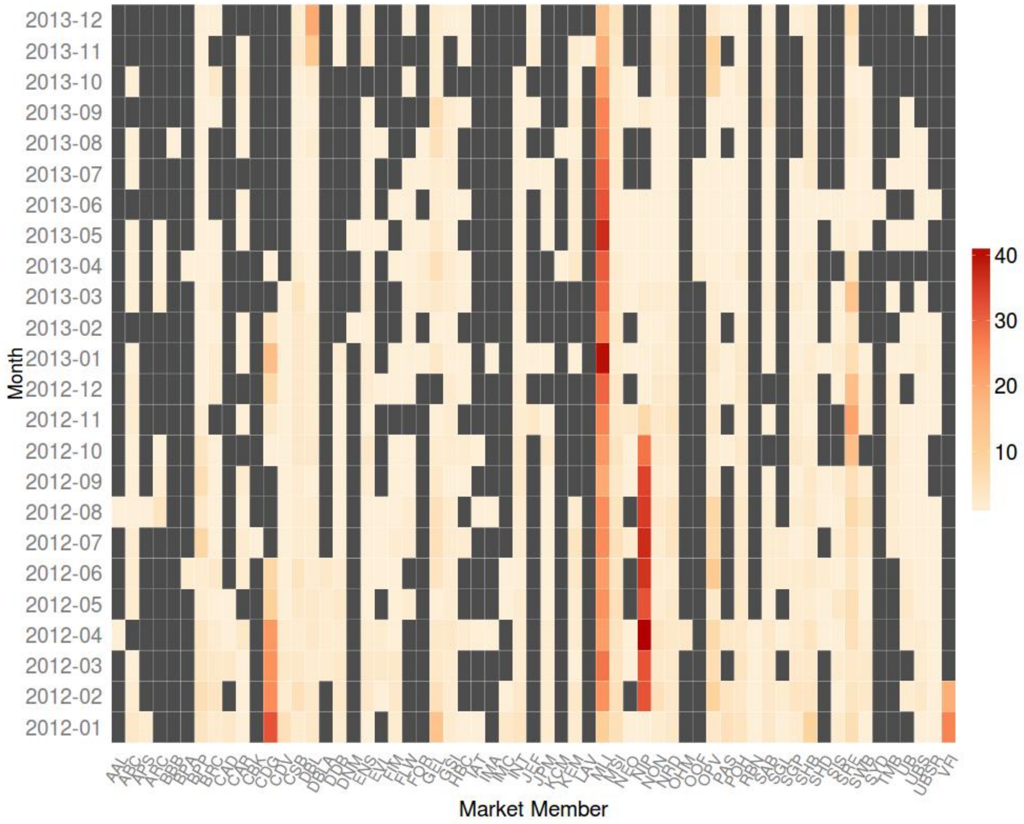
Degree of MMs
(under-expressed edges)



HF trades within a time window of w=1 msec
Similar results are obtained for w=50 msec

Nokia

Degree of MMs
(over-expressed edges)



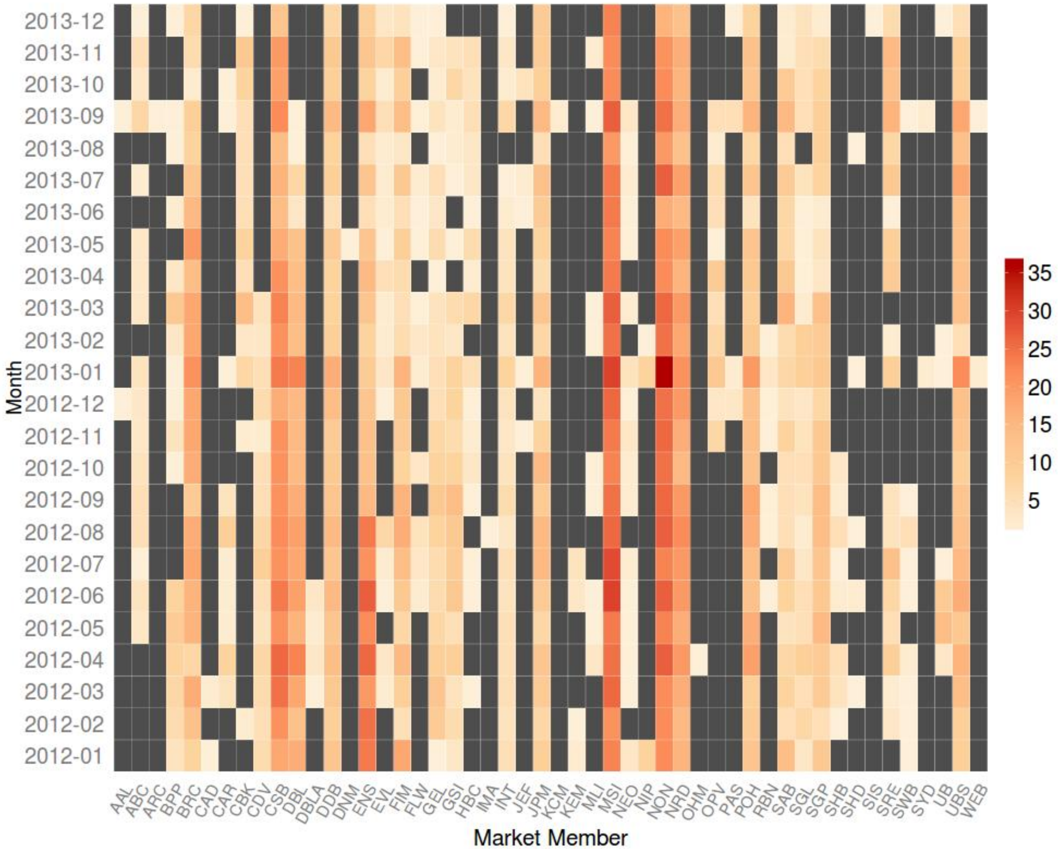
HF trades with *w*=1 msec

MM	min degree	mean degree	max degree
NIP	0	13.5	41
MLI	10	25.6	40
CDG	0	6.42	32
VFI	0	1.87	26
SRE	1	6.00	21
DBL	1	3.08	20
GEL	0	3.42	14
OPV	1	4.08	12
SHB	1	2.58	9
BPP	1	2.42	8
...	...	...	...
...	...	...	...

NIP	Nomura International plc
MLI	Merrill Lynch International
CDG	Citadel Securities (Europe) Limited
VFI	Virtu Financial Ireland Ltd
SRE	Spire Europe Limited
DBL	Deutsche Bank AG
GEL	KCG Europe Limited
OPV	Optiver VOF
SHB	Svenska Handelsbanken AB
BPP	BNP Paribas Arbitrage SNC
...	...

Nokia

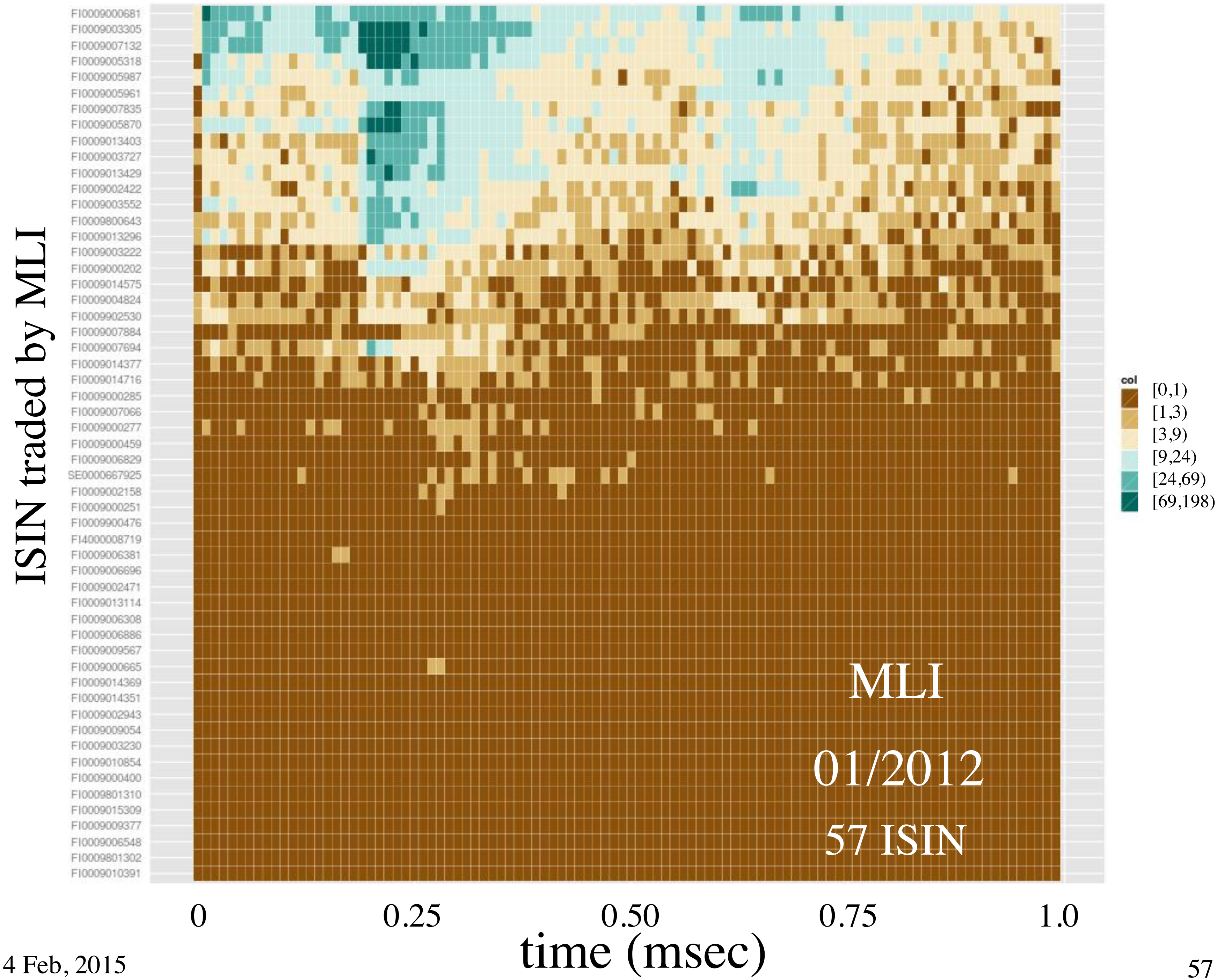
Degree of MMs (under-expressed edges)



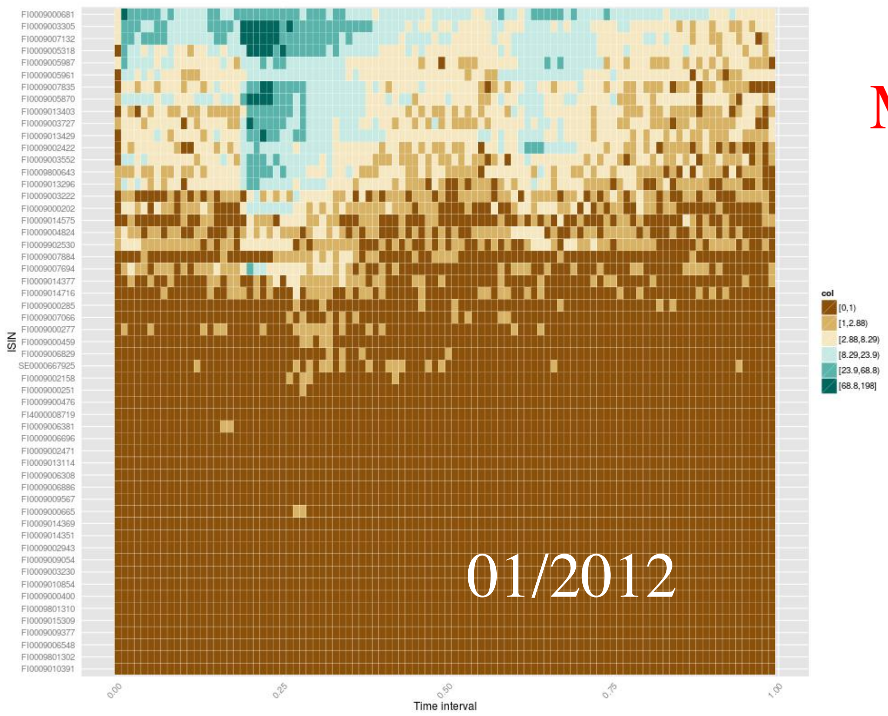
MM	min degree	mean degree	max degree
NON	16	23.9	36
MSI	20	24.7	30
ENS	5	15.3	27
CSB	12	20.0	26
DBL	0	12.3	23
NRD	13	17.6	22
UBS	5	12.5	22
BRC	2	13.5	21
POH	4	12.7	20
FIM	3	10.1	18
...	...	...	...
...	...	...	...

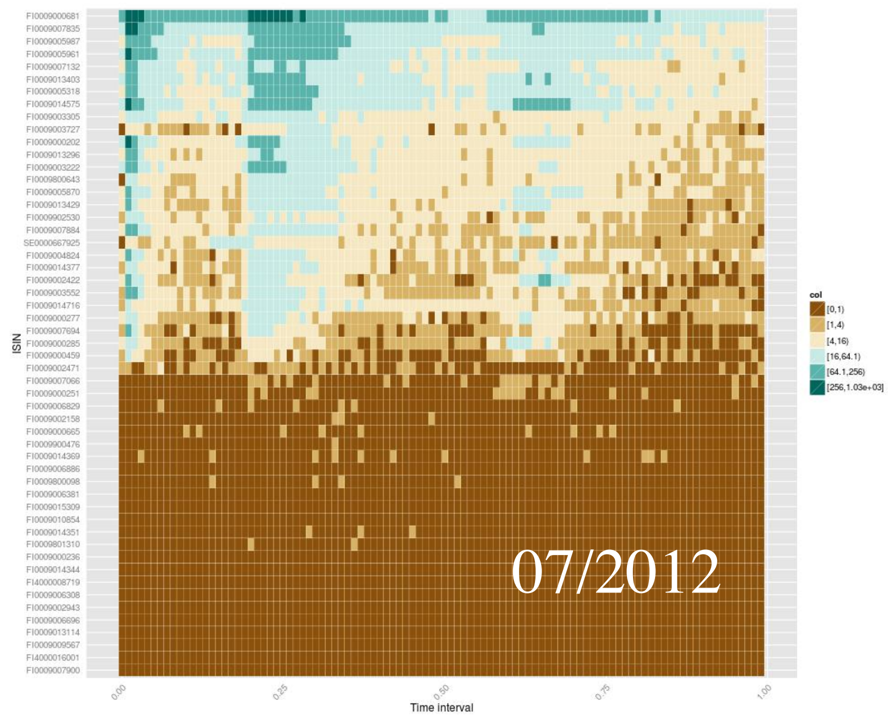
NON	Nordnet Bank AB
MSI	Morgan Stanley & Co. Int. Plc.
ENS	Skandinaviska Enskilda Banken AB
CSB	Credit Suisse Securities (Europe) Ltd
DBL	Deutsche Bank AG
NRD	Nordea Bank Finland Plc
UBS	UBS Limited
BRC	Barclays Capital Securities Ltd Plc
POH	Pohjola Bank Plc
FIM	FIM Bank Ltd.
...	...

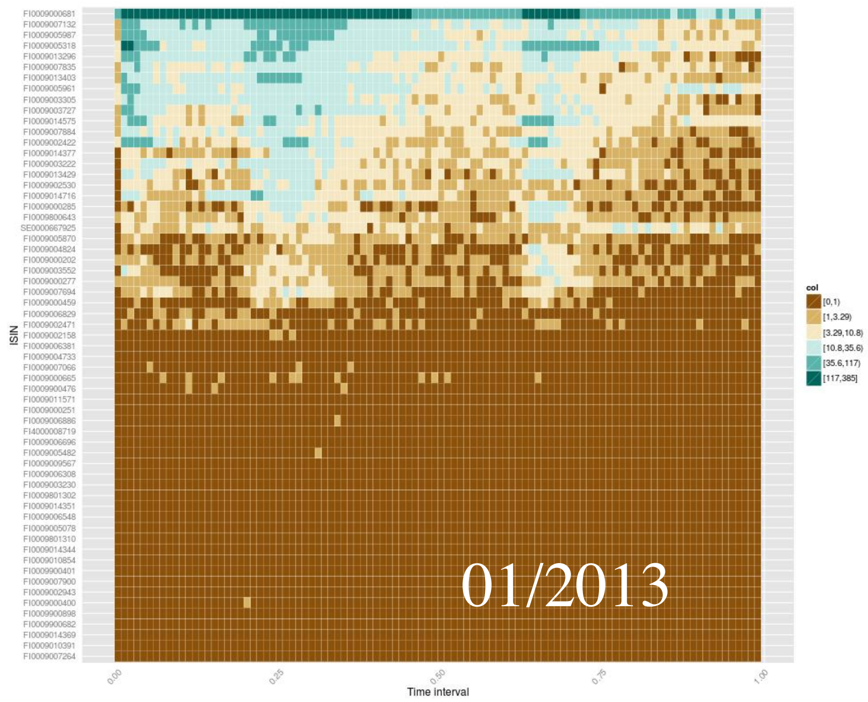
HF trades with *w*=1 msec

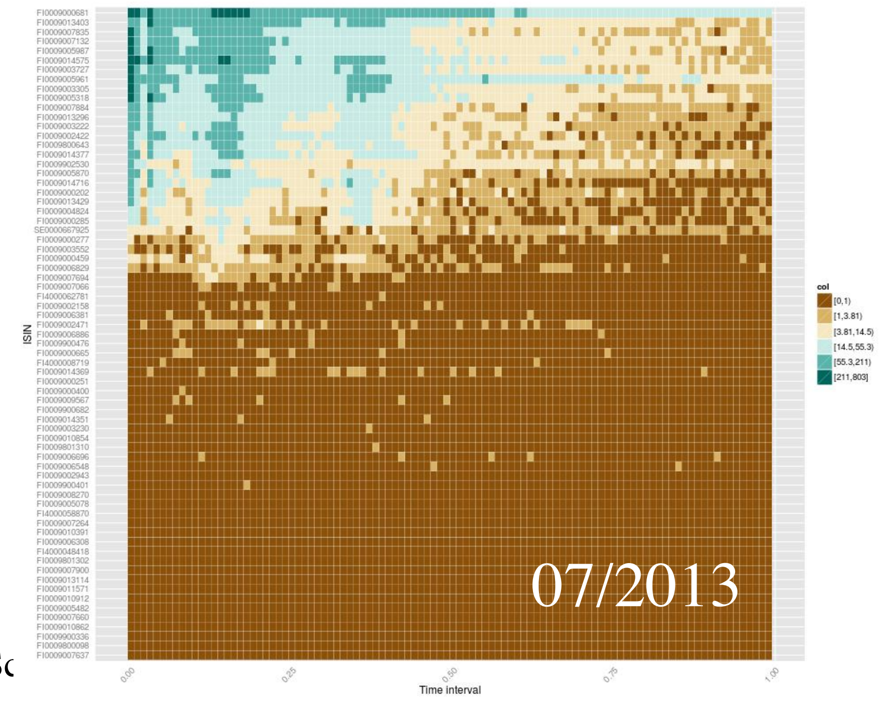

ISIN traded by MLI
FI0009000681
FI0009003305
FI0009007132
FI0009005318
FI0009005987
FI0009005961
FI0009007835
FI0009005870
FI0009013403
FI0009003727
FI0009013429
FI0009002422
FI0009003552
FI0009800643
FI0009013296
FI0009003222
FI0009000202
FI0009014575
FI0009004824
FI0009902530
FI0009007884
FI0009007694
FI0009014377
FI0009014716
FI0009000285
FI0009007066
FI0009000277
FI0009000459
FI0009006829
SE0000667925
FI0009002158
FI0009000251
FI0009900476
FI4000008719
FI0009006381
FI0009006696
FI0009002471
FI0009013114
FI0009006308
FI0009006886
FI0009009567
FI0009000665
FI0009014369
FI0009014351
FI0009002943
FI0009009054
FI0009003230
FI0009010854
FI0009000400
FI0009801310
FI0009015309
FI0009009377
FI0009006548
FI0009801302
FI0009010391
col
[0,1)
[1,3)
[3,9)
[9,24)
[24,69)
[69,198)
MLI
01/2012
57 ISIN
0
0.25
0.50
0.75
1.0
time (msec)

MLI


01/2012
ISIN
Time interval
0.00
0.25
0.50
0.75
1.00
col
[0,1)
[1,2.88)
[2.88,8.29)
[8.29,23.9)
[23.9,68.8)
[68.8,198]


07/2012
ISIN
Time interval
0.00
0.25
0.50
0.75
1.00
col
[0,1)
[1,4)
[4,16)
[16,64.1)
[64.1,256)
[256,1.03e+03]


01/2013
ISIN
Time interval
0.00
0.25
0.50
0.75
1.00
col
[0,1)
[1,3.29)
[3.29,10.8)
[10.8,35.6)
[35.6,117)
[117,385]


07/2013
ISIN
Time interval
0.00
0.25
0.50
0.75
1.00
col
[0,1)
[1,3.81)
[3.81,14.5)
[14.5,55.3)
[55.3,211)
[211,803]

Conclusions

- Proximity based networks are quite informative in finance;
- Different types of "association networks" can highlight different information;
- Statistically validated networks are able to detect over-expression and under-expression of events or relationships and can be useful to highlight the presence of a networked structure of markets in heterogeneous complex systems.