Similarity-based & Statistically Validated Networks in Finance

Rosario Nunzio Mantegna
Central European University, Budapest, Hungary
Palermo University, Palermo, Italy
Outline

• I will discuss the concept of similarity based networks and their use in finance;

• I will present the methodology of statistically validated networks by discussing its application to syndicated loans and the interbank market.
Two different approaches in building networks

**Event or relation defined networks**

Example: nodes are banks, links are credit relationships

**Similarity-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.
The first investigation of a correlation based networks


Fig. 1. (a) Minimal spanning tree connecting the 30 stocks used to compute the Dow Jones Industrial Average. The 30 stocks are labeled by their tick symbols. The distance between the stocks is bounded as: CHV-TX 0.90 < d(i, j) ≤ 0.95; XON-TX 0.95 < d(i, j) ≤ 1.00; KO-PG 1.00 < d(i, j) ≤ 1.05; MMM-GE-KO, DD-GE-T, AA-IP and MRK-KO-MCD 1.05 < d(i, j) ≤ 1.10; CAT-IP-MMM, AXP-JPM-GE-GM, BA-GE-UTX, DD-XON and MO-PG 1.10 < d(i, j) ≤ 1.15; DIS-GE-EK, DD-UK, BS-IP-ALD and GE-WX 1.15 < d(i, j) ≤ 1.20; AA-GT, GE-IBM, KO-Z and IP-S 1.20 < d(i, j) ≤ 1.25. (b) Hierarchical tree of the subdominant ultrametric space associated with the minimal spanning tree of a). In the hierarchical tree, several groups of stocks homogeneous with respect to the economic activities of the companies are detected: (i) oil companies (Exxon (XON), Texaco (TX) and Chevron (CHV)); (ii) raw material companies (Alcoa (AA) and International paper (IP)) and (iii) companies working in the sectors of consumer nondurable products (Procter & Gamble (PG)) and food and drinks (Coca Cola (KO)). The ultrametric distance at which individual stocks are branching from the tree is given by the y axis.
Filtering the correlation matrix using single linkage clustering

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

\[ d_{ij} = \sqrt{2(1 - \rho_{ij})} \]

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

- AXP  MER  0.664  0.820
- IBM  MER  0.617  0.875
- SLB  OXY  0.592  0.903
- BAC  MER  0.591  0.904
- RD   OXY  0.590  0.905
- TXN  MOT  0.582  0.914
- IBM  TXN  0.552  0.947
- AXP  BAC  0.547  0.952
- AIG  AXP  0.543  0.956
- AXP  IBM  0.537  0.962
- SLB  RD   0.533  0.966
- MER  TXN  0.533  0.966
- AIG  MER  0.529  0.970
- AIG  BAC  0.518  0.982
- IBM  MOT  0.475  1.025
- MOT  MER  0.462  1.037
- MER  RD   0.440  1.058
- AXP  TXN  0.422  1.075

23 Sep 2014
The hierarchical tree obtained from single linkage clustering algorithm has information equivalent to a simplified matrix having only n-1 distinct elements. It can be proven that such a matrix is an ultrametric matrix when a distance is defined between each pair of elements.

<table>
<thead>
<tr>
<th></th>
<th>AIG</th>
<th>IBM</th>
<th>BAC</th>
<th>AXP</th>
<th>MER</th>
<th>TXN</th>
<th>SLB</th>
<th>MOT</th>
<th>RD</th>
<th>OXY</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIG</td>
<td>1</td>
<td>0.543</td>
<td>0.543</td>
<td>0.543</td>
<td>0.543</td>
<td>0.440</td>
<td>0.543</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
</tr>
<tr>
<td>IBM</td>
<td>1</td>
<td>0.591</td>
<td>0.617</td>
<td>0.617</td>
<td>0.552</td>
<td>0.440</td>
<td>0.552</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
</tr>
<tr>
<td>BAC</td>
<td>1</td>
<td>0.591</td>
<td>0.591</td>
<td>0.552</td>
<td>0.440</td>
<td>0.552</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
</tr>
<tr>
<td>AXP</td>
<td>1</td>
<td>0.664</td>
<td>0.552</td>
<td>0.440</td>
<td>0.552</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
</tr>
<tr>
<td>MER</td>
<td>1</td>
<td>0.552</td>
<td>0.440</td>
<td>0.582</td>
<td>0.440</td>
<td>0.440</td>
<td>0.590</td>
<td>0.592</td>
<td>0.440</td>
<td>0.440</td>
</tr>
<tr>
<td>TXN</td>
<td>1</td>
<td>0.440</td>
<td>0.590</td>
<td>0.440</td>
<td>0.440</td>
<td>0.592</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
</tr>
<tr>
<td>SLB</td>
<td>1</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
<td>0.592</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
</tr>
<tr>
<td>MOT</td>
<td>1</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
<td>0.592</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
</tr>
<tr>
<td>RD</td>
<td>1</td>
<td>0.590</td>
<td>0.590</td>
<td>0.590</td>
<td>0.590</td>
<td>0.592</td>
<td>0.590</td>
<td>0.590</td>
<td>0.590</td>
<td>0.590</td>
</tr>
<tr>
<td>OXY</td>
<td>1</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
<td>0.592</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
<td>0.440</td>
</tr>
</tbody>
</table>

\[ C_{\text{SL}} \]
Kruskal's algorithm of the Minimum Spanning Tree

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

\[
\begin{align*}
&\text{C}_{\text{SL}} \\
&\begin{array}{c}
\text{MOT} \\
\text{SLB} \\
\text{IBM} \\
\text{TXN} \\
\text{BAC} \\
\text{AXP} \\
\text{MER} \\
\text{OXY} \\
\text{RD}
\end{array}
\end{align*}
\]

\[
\begin{array}{c|c|c|c|c|c|c|c|c|c|c}
& \text{AIG} & \text{IBM} & \text{BAC} & \text{AXP} & \text{MER} & \text{TXN} & \text{SLB} & \text{MOT} & \text{RD} & \text{OXY} \\
\hline
\text{AIG} & 1 & 0.413 & 0.518 & 0.543 & 0.529 & 0.341 & 0.271 & 0.231 & 0.412 & 0.294 \\
\text{IBM} & & 1 & 0.471 & 0.537 & 0.617 & 0.552 & 0.298 & 0.475 & 0.373 & 0.270 \\
\text{BAC} & & & 1 & 0.547 & 0.591 & 0.400 & 0.258 & 0.349 & 0.370 & 0.276 \\
\text{AXP} & & & & 1 & 0.664 & 0.422 & 0.347 & 0.351 & 0.414 & 0.269 \\
\text{MER} & & & & & 1 & 0.533 & 0.344 & 0.462 & 0.440 & 0.318 \\
\text{TXN} & & & & & & 1 & 0.305 & 0.582 & 0.355 & 0.245 \\
\text{SLB} & & & & & & & 1 & 0.193 & 0.533 & 0.592 \\
\text{MOT} & & & & & & & & 1 & 0.258 & 0.166 \\
\text{RD} & & & & & & & & & 1 & 0.590 \\
\text{OXY} & & & & & & & & & & 1
\end{array}
\]
The filtering selects part of the information of a distance matrix

\[ d_{ij} = \sqrt{2(1 - \rho_{ij})} \]

100 highly capitalized stocks of US equity markets 1995-1998 daily data
The minimum spanning tree is the most basic and robust way to obtain a similarity based network. There are many generalizations of this basic approach. A prominent one is the Planar Maximally Filtered Graph.

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 Spanning 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

Planar graphs

A graph is **planar** if its edges can be embedded on a surface of genus 0, i.e. a surface like a plane or a sphere, without intersections of the edges.

N=100 (US markets) daily returns
1995-1998 T=1011
Partial correlation network

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.

We investigated the quantity

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

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.

---

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

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

Specifically, we investigate

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

We use this directed similarity measure to obtain a Partial Correlation Planar Graph
The Partial Correlation Planar Graph (economic subsectors)
A statistical assessment of links of similarity based networks can be performed by using bootstrap replicas.

Data Set

<table>
<thead>
<tr>
<th>t₁</th>
<th>V₁</th>
<th>V₂</th>
<th>V₃</th>
<th>...</th>
<th>Vₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.113</td>
<td>1.123</td>
<td>-0.002</td>
<td>...</td>
<td>0.198</td>
<td></td>
</tr>
</tbody>
</table>

| t₂  | 1.567 | 0.789 | 0.842 | ... | -0.234 |

| t₃  | 1.065 | -1.962 | 0.567 | ... | 1.785 |

| t₄  | 1.112 | 0.998 | -0.424 | ... | 2.735 |

| t₅  | -0.211 | 0.312 | -0.0217 | ... | 0.587 |

| ... | ... | ... | ... | ... | ... |

| T   | 0.479 | -1.828 | -2.041 | ... | -0.193 |

Pseudo-replicate Data Set

<table>
<thead>
<tr>
<th>V₁</th>
<th>V₂</th>
<th>V₃</th>
<th>...</th>
<th>Vₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.567</td>
<td>0.789</td>
<td>0.842</td>
<td>...</td>
<td>-0.234</td>
</tr>
</tbody>
</table>

| 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.
Statistical reliability of the minimum spanning tree


Financial Risk & Network Theory - Cambridge
Bootstrap vs correlation

For Gaussian series: \( \sigma_\rho = \frac{1 - \rho^2}{\sqrt{T - 3}} \)

\( N_v = 300 \)
Stocks of US equity markets
2001-2003
\( T = 748 \)
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?

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

Example:

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 $N_A$ the number of packages that bank $A$ has subscribed and $N_B$ 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.

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:

\[
P(X \mid N, N_A, N_B) = \binom{N_A}{X} \binom{N - N_A}{N_B - X} \binom{N}{N_B}
\]

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

p-value associated to a detection of
coco-occurrence ≥ X:

\[
p = 1 - \sum_{i=0}^{X-1} \binom{N_A}{i} \binom{N - N_A}{N_B - i} \binom{N}{N_B}
\]
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 $\theta = 0.01/T$ where $T$ is the number of tests to be done.

Another type of correction (less restrictive) is the so-called False Discovery Rate correction.
Network of banks performing syndicated loans

- **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

Year 2012

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.

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

$\text{OVER-expression:}$

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

$\text{UNDER-expression:}$
By using this approach we have shown that the e-MID market presents statistically validated links.

A simple model of dynamic network of the interbank market showing statistically validated links.

At each transaction a lender and a borrower that aim to do a transaction are selected. The probability that the selected lender accept the selected borrower is proportional to an attractiveness $w$ common to all borrower and to a trust proxy of the specific borrower obtained by considering the number of past credit relationships commonly undertaken. The trust is built within a memory time interval which is a parameter of the model.

By calibrating the simulations of this model on the e-MID data we obtains\(^{\natural}\)

Conclusions

• Similarity based networks are quite informative in finance;

• Different networks can highlight different information;

• Statistically validated networks are able to detect over-occurrence and under-occurrence of events or relationships and can be useful to highlight the presence of a networked structure of markets.
I acknowledge funding from

Institute for New Economic Thinking

CRISIS
Complexity Research Initiative for Systemic Instabilities