Statistically-validated networks

Chester Curme

In collaboration with:
Michele Tumminello (University of Palermo)
Rosario N. Mantegna (University of Palermo)
H. Eugene Stanley (BU)
Dror Y. Kenett (BU)
Outline

• Why study networks?
• How can we construct networks from data?
  – Advantages and disadvantages of existing methods
• Statistically-validated network methodology
• Application to financial markets

http://arxiv.org/abs/1401.0462
Why study networks?

- Useful framework for the analysis of *complex systems*. 

? 
Why study networks?

• Useful framework for the analysis of complex systems.
How are networks built from data?

• Consider $N$ interacting units.
• Construct $N \times N$ matrix $C$, where $C_{ij}$ is a measure of similarity between units $i$ and $j$.
  – Common example is Pearson correlation:
    \[ C_{ij} = \rho_{ij} \]
• Filter elements of $C$ to edges of a network. Most methods for this step fall into two categories:
  – Topological/hierarchical methods
  – Threshold methods
Hierarchical Method: Minimal Spanning Tree

Mantegna 1999
Topological/hierarchical methods

• Advantages
  – Useful for obtaining “skeleton” or outline of important relationships in a system.
  – Intrinsically hierarchical.

• Disadvantages
  – Imposes topological constraints (e.g., tree structure with $N-1$ edges for MST).
  – No information about statistical significance of or uncertainty in the measures $C_{ij}$.
Threshold methods

• Construct network from edges $C_{ij} > C_c$, for some similarity threshold $C_c$.
• How to choose $C_c$?
Select threshold by statistical confidence?
Threshold methods

• Advantages
  – Robust to statistical uncertainty.
  – No topological constraints.

• Disadvantages
  – Difficult to find single appropriate threshold for all $C_{ij}$ that displays a hierarchical structure.
  – Fails to take into account heterogeneities in relationships among nodes.
Statistically-validated networks
Different thresholds for each pair.
• Account for heterogeneities among nodes.
• Associate $p$-value to each entry of $C$; construct network using edges with a $p$-value below a threshold.
• Illustrate method through concrete example: lagged correlations among returns of $N = 100$ stocks in NYSE.


• “Signals” are returns: $r_t = \ln(p_t) - \ln(p_{t-\Delta t})$.

• “Similarity measure” is lagged Pearson correlation.
What is Pearson Correlation?

\[ \tilde{x} \equiv \frac{x - \langle x \rangle}{\sqrt{\sum_t (x_t - \langle x \rangle)^2}} \]

\[ \rho_{xy} = \tilde{x} \cdot \tilde{y} \]
Construct empirical lagged correlation matrix $C_{ij}$

$\Delta t = 130$ min:
• How many shufflings do we need to perform to validate links with $p = 0.01$?
  – If we were just interested in one pair of stocks, we would need 100.
  – Because we are testing $N^2 = 100^2 = 10^4$ hypotheses, however, we need to perform at least $10^6$ shufflings to account for multiple comparisons.

• Two protocols for multiple comparisons: Bonferroni (conservative) and FDR (less conservative).
Distribution of lagged correlation coefficients for all pairs of $N = 100$ stocks at $\Delta t = 15$ min. Bounds of coefficients selected using both Bonferroni and FDR filtering procedures are shown.
Distribution of lagged correlation coefficients for all pairs of $N = 100$ stocks at $\Delta t = 15$ min. Bounds of coefficients selected using both Bonferroni and FDR filtering procedures are shown.
Network formed using returns sampled at $\Delta t = 130$ min. Transaction data are from 2011-2012.
Network formed using returns sampled at $\Delta t = 65$ min. Transaction data are from 2011-2012.
Network formed using returns sampled at $\Delta t = 30$ min. 
Transaction data are from 2011-2012.
Network formed using returns sampled at $\Delta t = 15$ min.
Transaction data are from 2011-2012.
Network formed using returns sampled at $\Delta t = 5$ min.
Transaction data are from 2011-2012.
Market efficiency: leaders

Out-degree distributions (FDR networks)
Market efficiency: followers

In-degree distributions (FDR networks)
Comparison with threshold method

<table>
<thead>
<tr>
<th>Δt</th>
<th># pos. valid. (bootstrap)</th>
<th># pos. valid. (normal dist.)</th>
<th># pos. valid (both)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 min</td>
<td>2,252</td>
<td>2,398</td>
<td>2,230</td>
</tr>
<tr>
<td>15 min</td>
<td>681</td>
<td>754</td>
<td>666</td>
</tr>
<tr>
<td>30 min</td>
<td>134</td>
<td>158</td>
<td>131</td>
</tr>
<tr>
<td>65 min</td>
<td>29</td>
<td>43</td>
<td>26</td>
</tr>
<tr>
<td>130 min</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>
Ongoing work

• How might lagged correlations contribute to phenomena observed using synchronous correlations?
  – Economic sector clustering

• What changes occur during the trading day?
Hierarchical Method: Minimal Spanning Tree

Mantegna 1999
Intra-day seasonalities

$\Delta t = 130$ min:

Day 1

Stock $i$

Day 2

Stock $j$
What changes occur in the course of a trading day?

Correlations between returns in first 15 min. of trading day with returns in second 15 min. of trading day.
What changes occur in the course of a trading day?

Correlations between last two 15min. periods in the trading day.
Conclusions

• We have introduced a new tool for the analysis of complex systems.
• Statistically-validated networks are constructed without imposing any topology, and account for heterogeneities in relationships among nodes at the expense of computation time.
• The method is ideally suited to the analysis of lagged correlations in financial markets. We find:
  – Increase in network connectivity with decreasing time of return sampling $\Delta t$.
  – Increase in market efficiency over the past decade.
Future work

• Intra-day seasonalities
• Relation to Epps Effect
• Hayashi-Yoshida estimator
• Prediction model: to what extent are these relationships exploitable in the presence of market frictions?

Thank you!
References

Epps curves of all synchronous correlation coefficients.

Distributions of correlations at $\Delta t = 15$ min.
Effect of time series length on statistical power at a fixed time horizon $\Delta t = 15$ min.
Magnitude of filtered correlations

2002-2003

2011-2012
Effect of changing time lag at a fixed time horizon $\Delta t = 15$ min.
Out-degree distributions (FDR networks) normalized by total number of links
Market efficiency

In-degree distributions (FDR networks) normalized by total number of links