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)

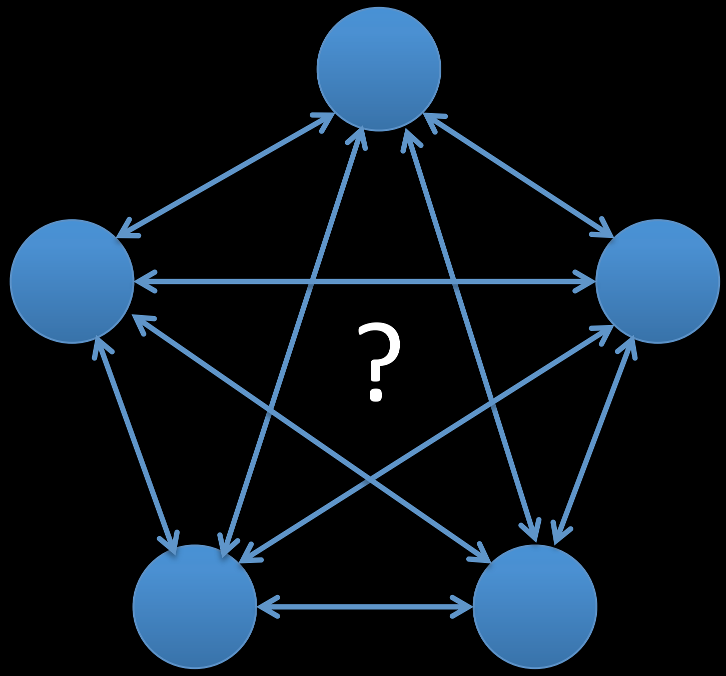
Outline

- 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

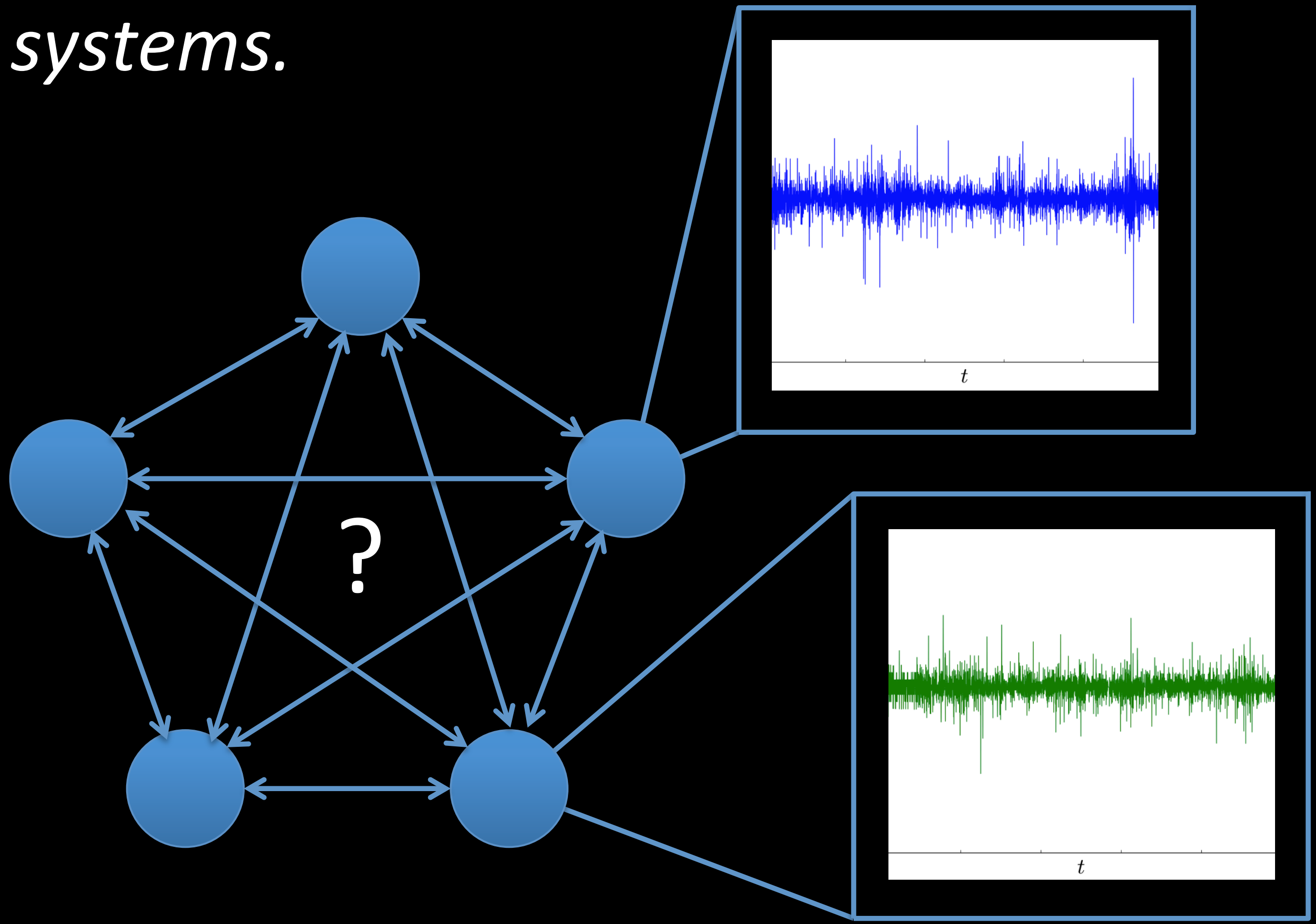
Why study networks?

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



Why study networks?

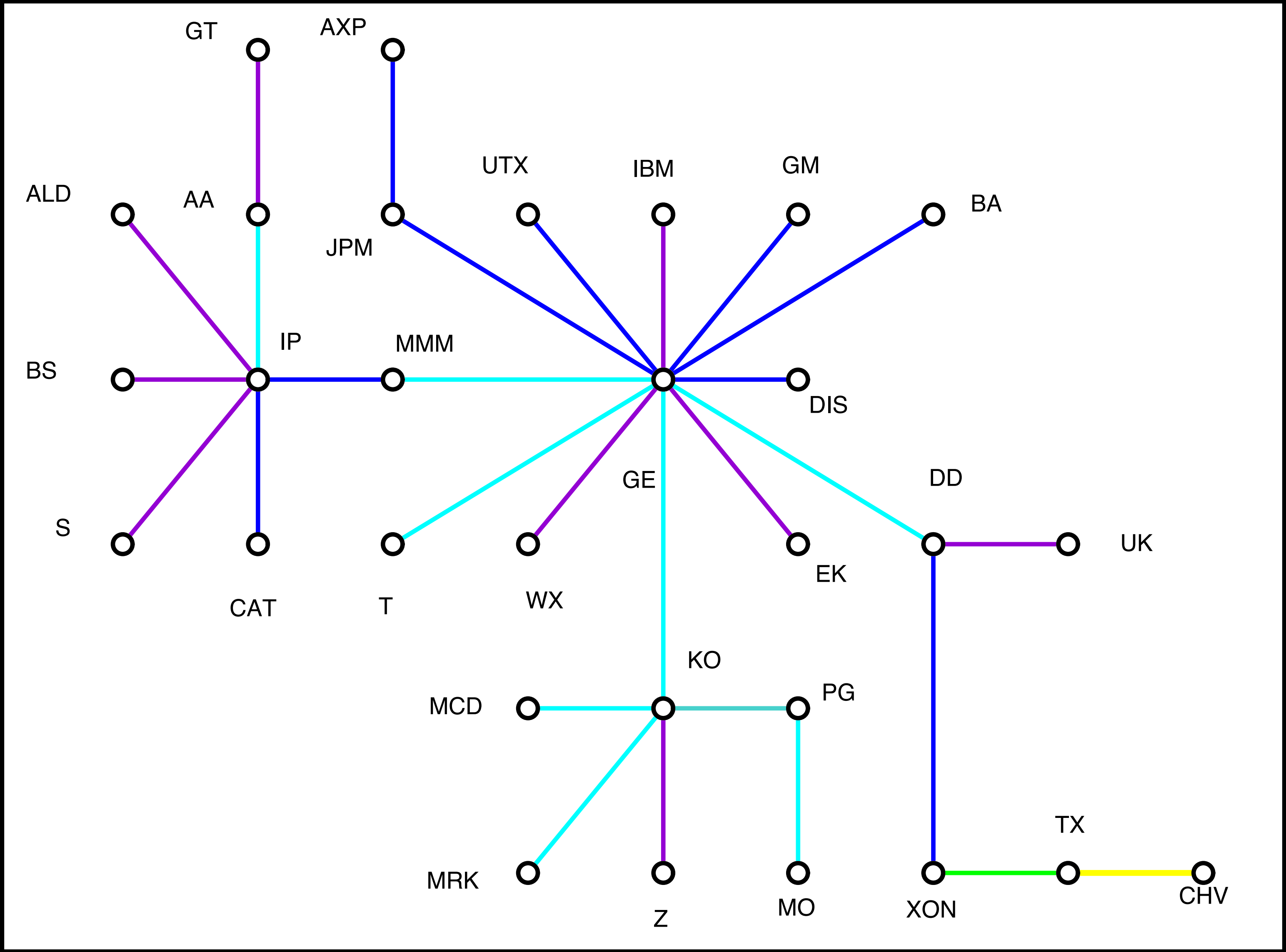
- Useful framework for the analysis of *complex systems.*



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

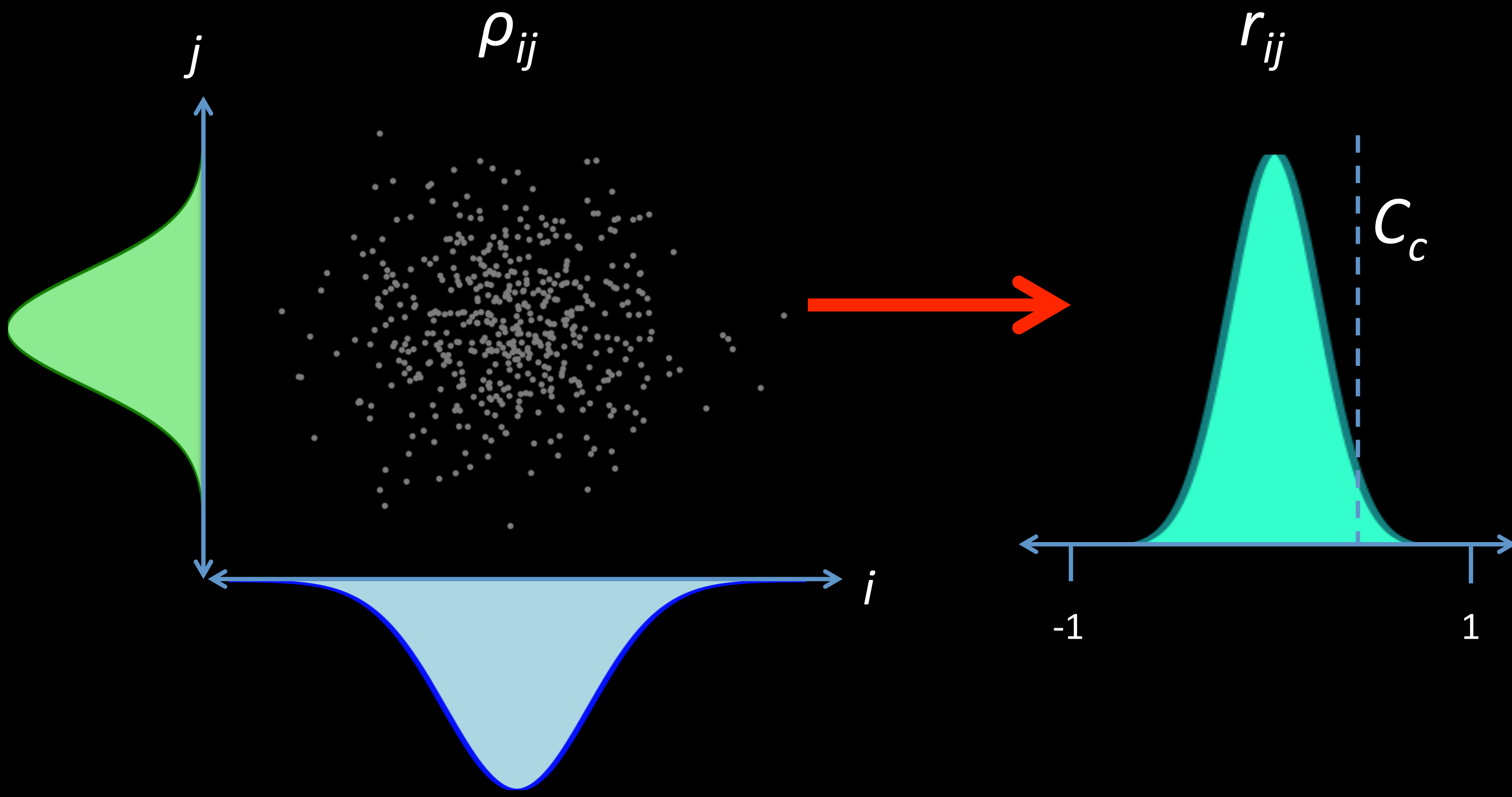
Topological/hierarchical methods

- 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?

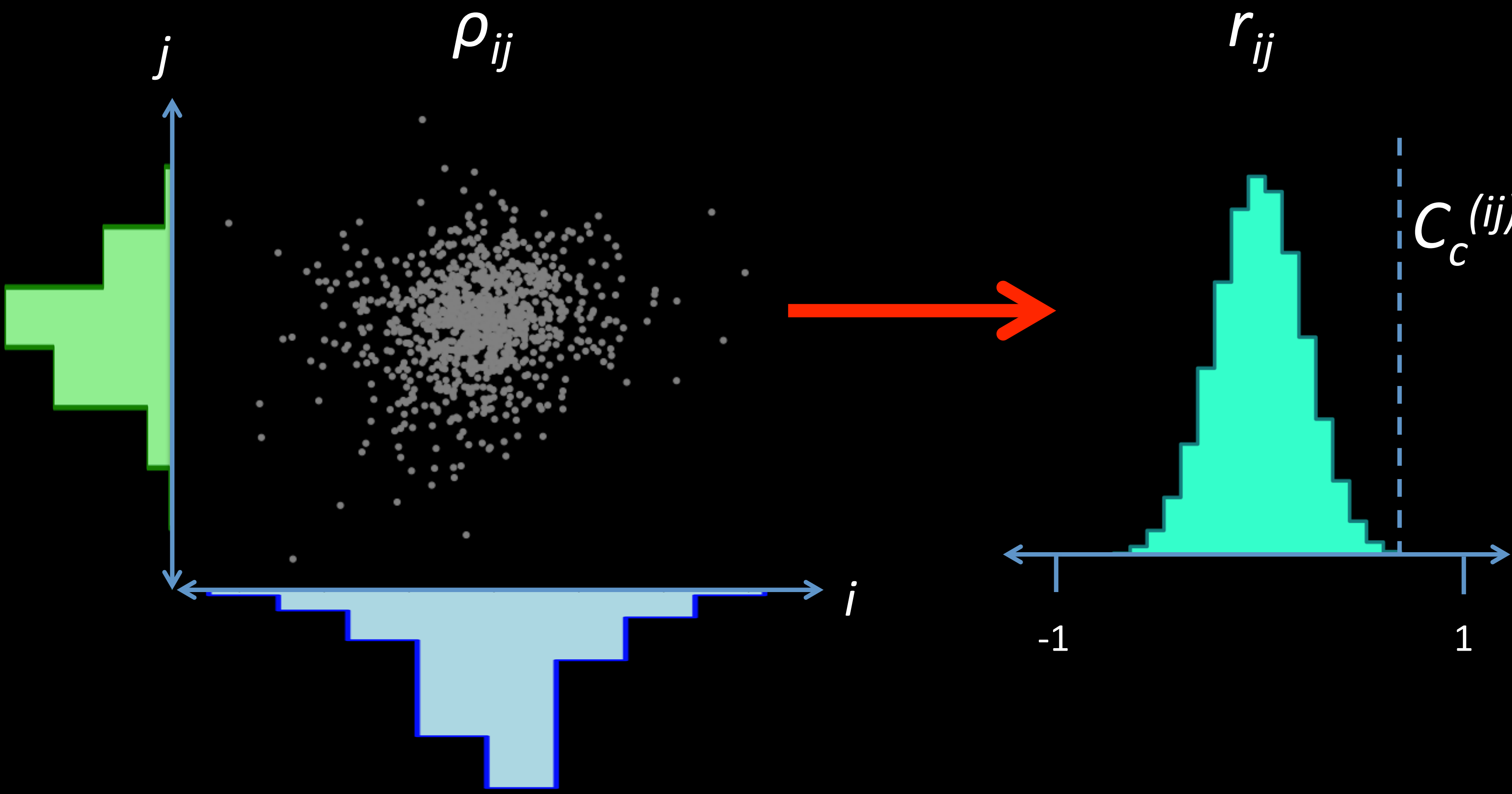


Threshold methods

- 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.
- Compare results from two datasets: 2002-2003 and 2011-2012.
- "Signals" are returns: $r_t = \ln(p_t) - \ln(p_{t-\Delta t})$.
- "Similarity measure" is lagged Pearson correlation.

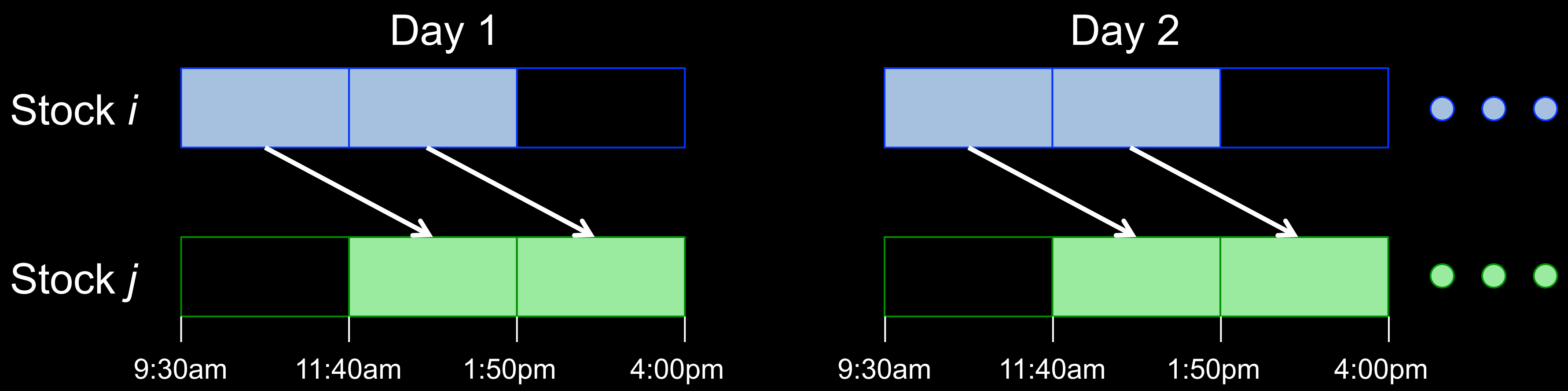
What is Pearson Correlation?

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

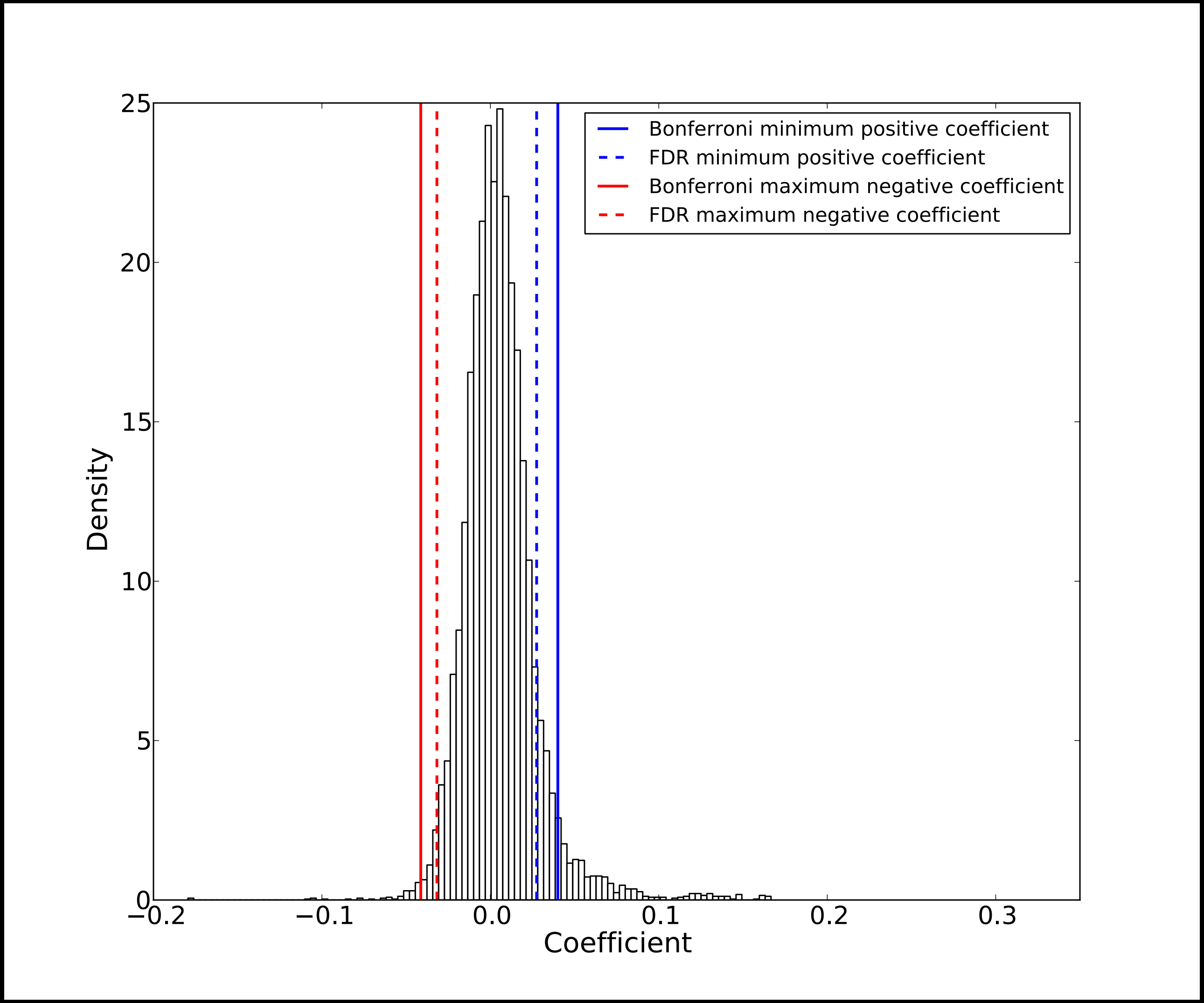
$$\rho_{xy} = \tilde{\vec{x}} \cdot \tilde{\vec{y}}$$

Construct empirical lagged correlation matrix C_{ij}

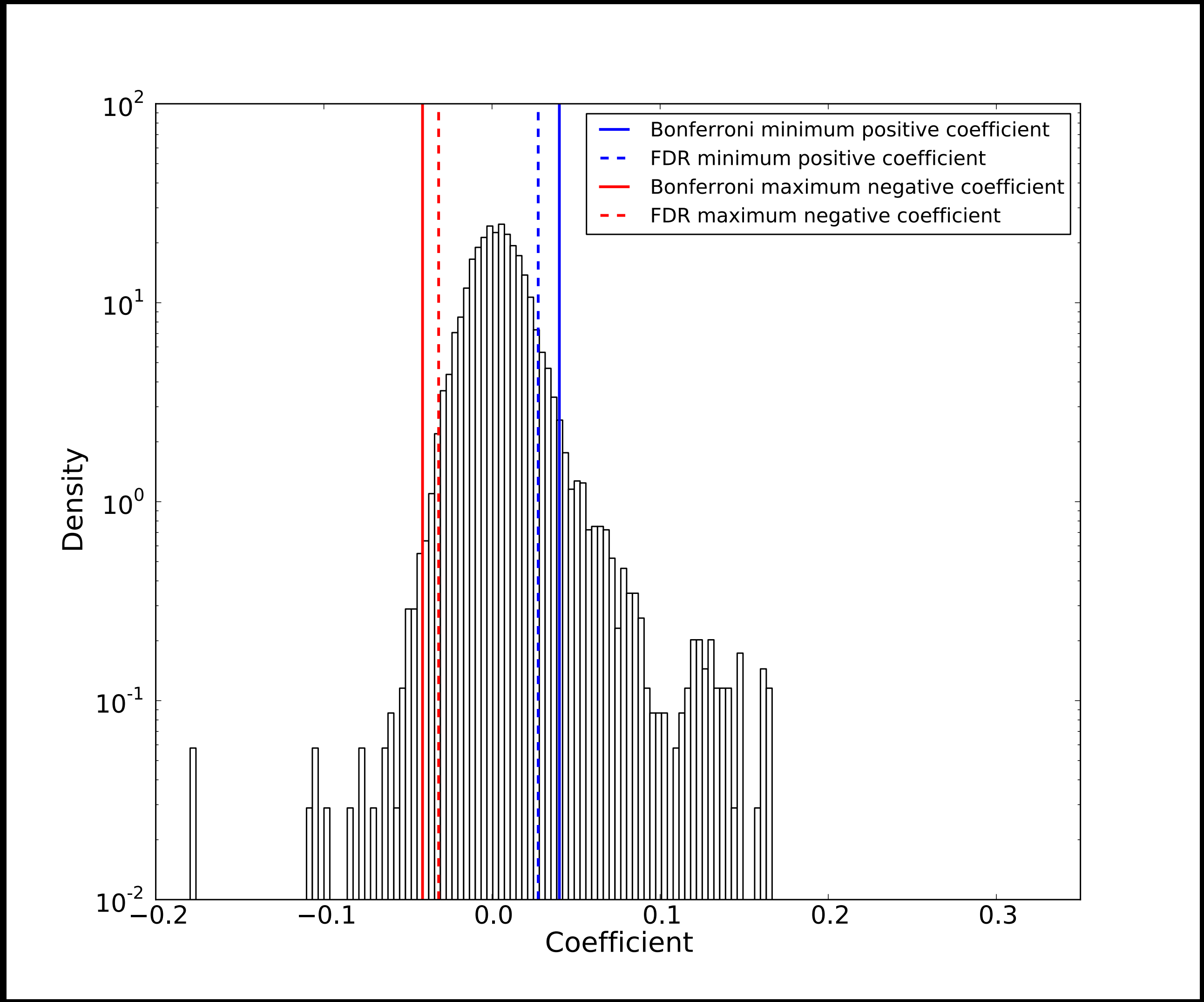
Δt = 130 min:



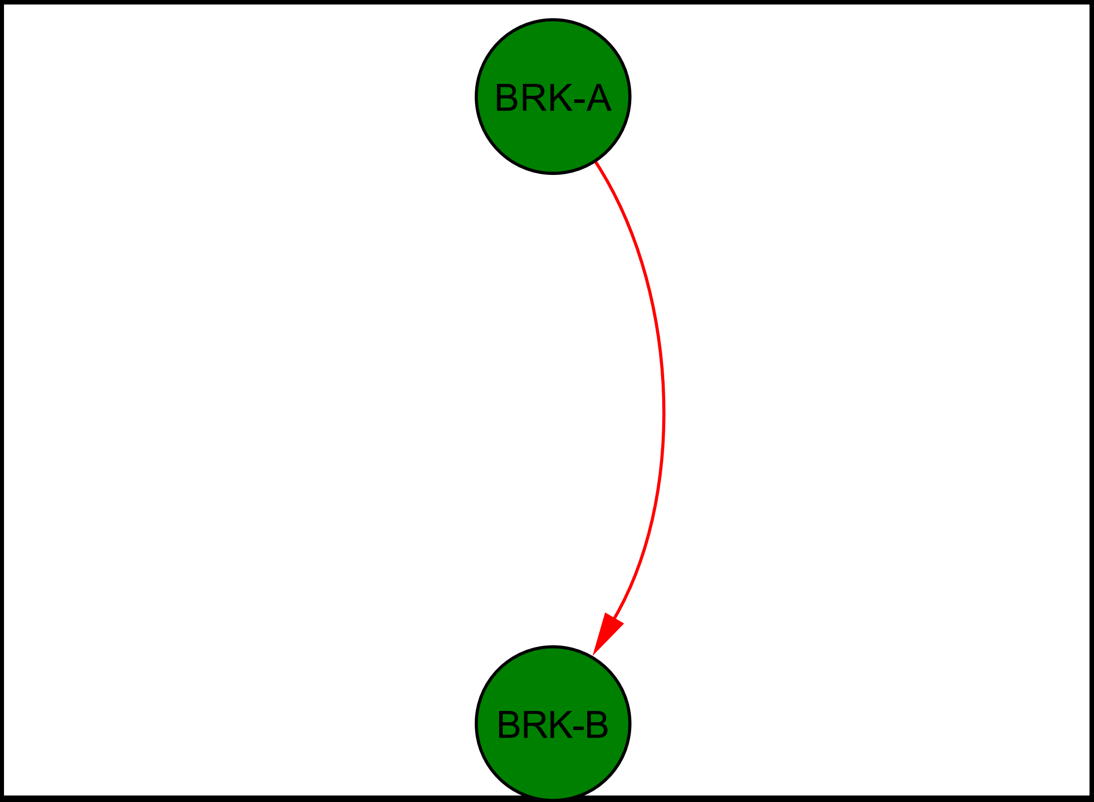
- How many shufflings do we need to perform to validate links with $p = 0.01$?
  - If we were just interested in one pair of stocks, we would need 100.
  - Because we are testing $N^2 = 100^2 = 10^4$ hypotheses, however, we need to perform at least 10^6 shufflings to account for multiple comparisons.
- Two protocols for multiple comparisons: Bonferroni (conservative) and FDR (less conservative).



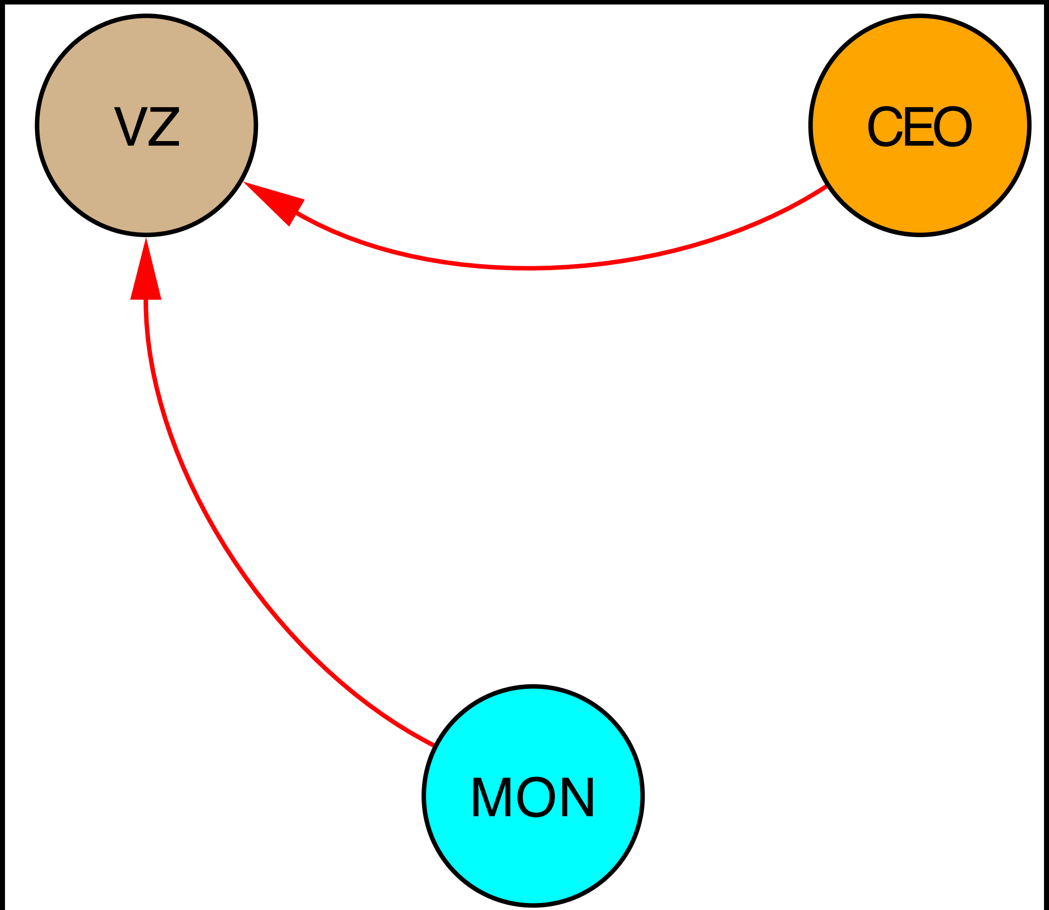
Distribution of lagged correlation coefficients for all pairs of N = 100 stocks at Δt = 15 min. Bounds of coefficients selected using both Bonferroni and FDR filtering procedures are shown.



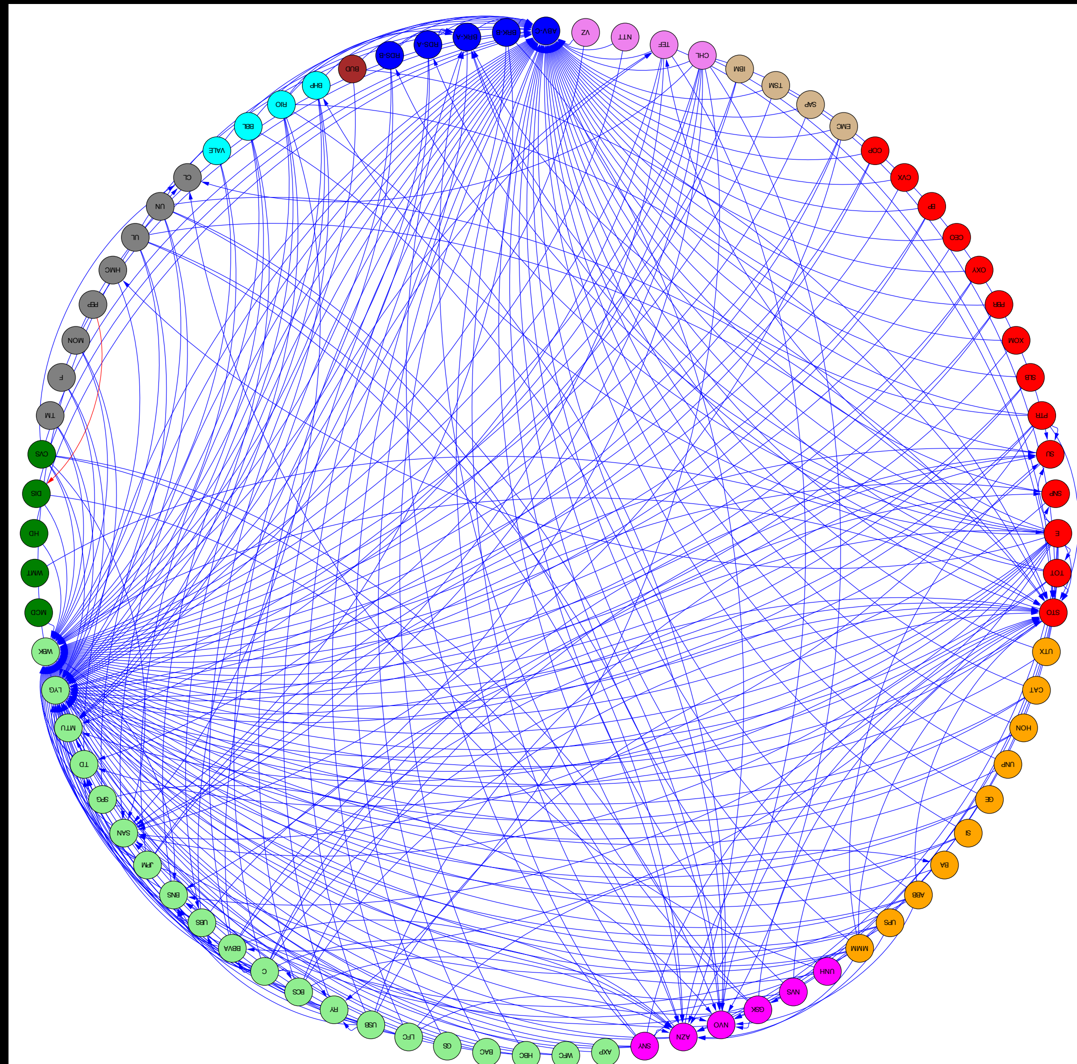
Distribution of lagged correlation coefficients for all pairs of N = 100 stocks at Δt = 15 min. Bounds of coefficients selected using both Bonferroni and FDR filtering procedures are shown.



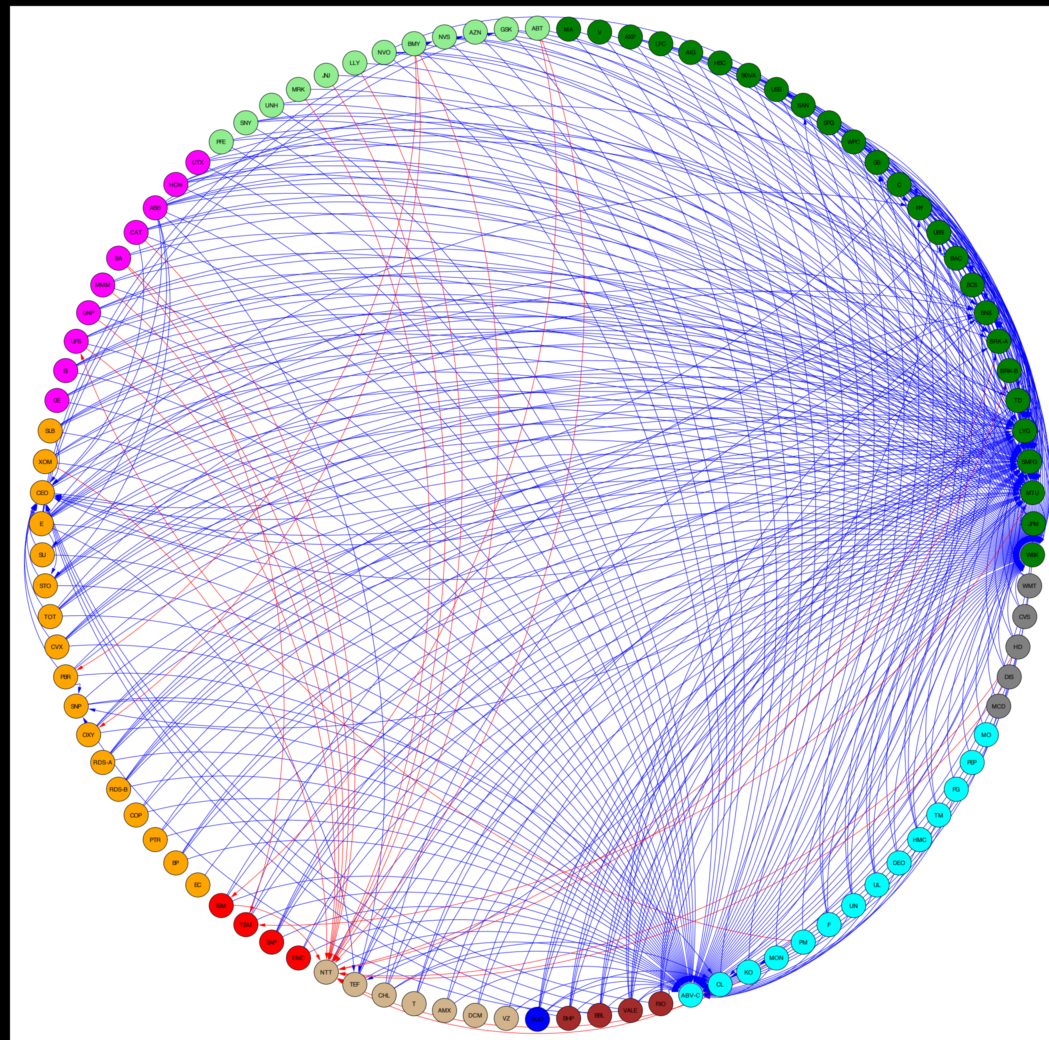
Network formed using returns sampled at Δt = 130 min.
Transaction data are from 2011-2012.



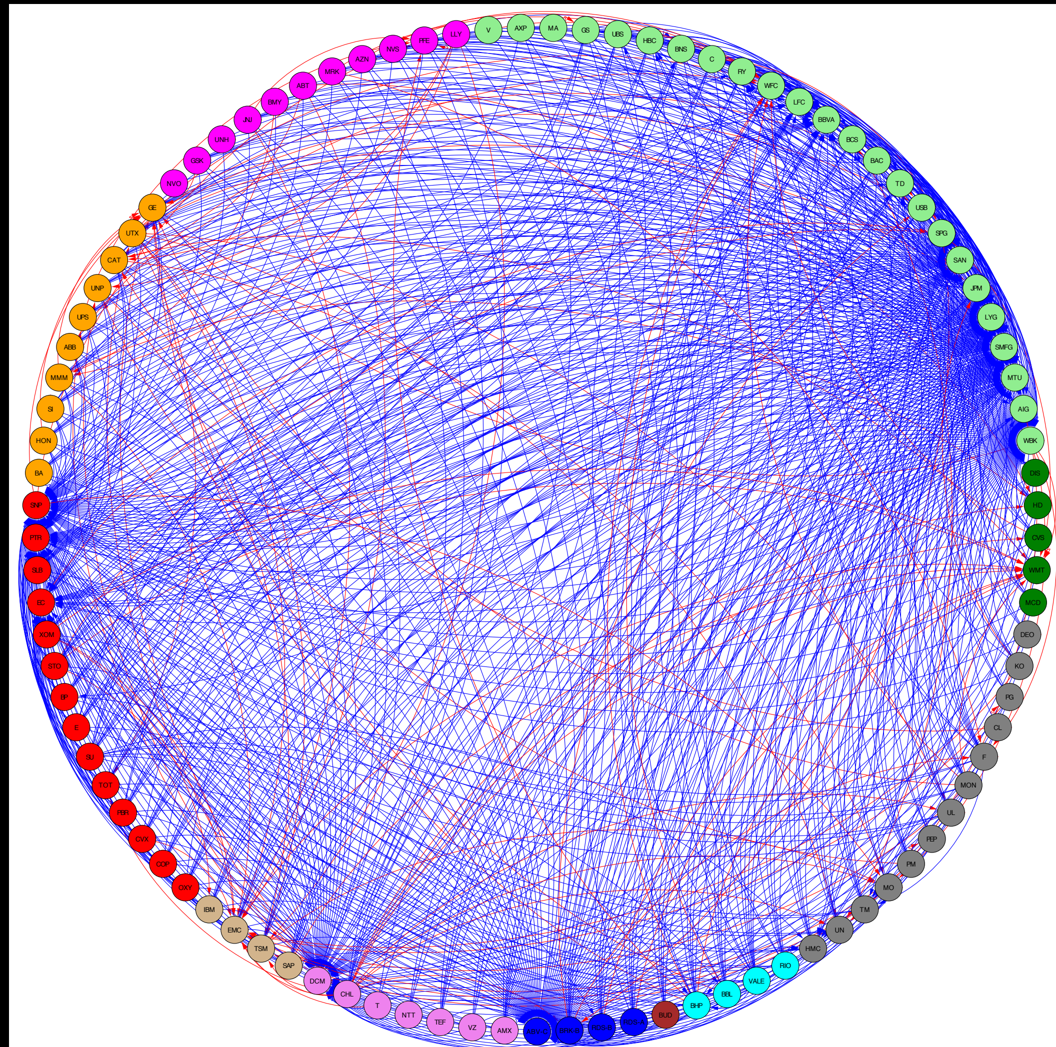
Network formed using returns sampled at Δt = 65 min.
Transaction data are from 2011-2012.



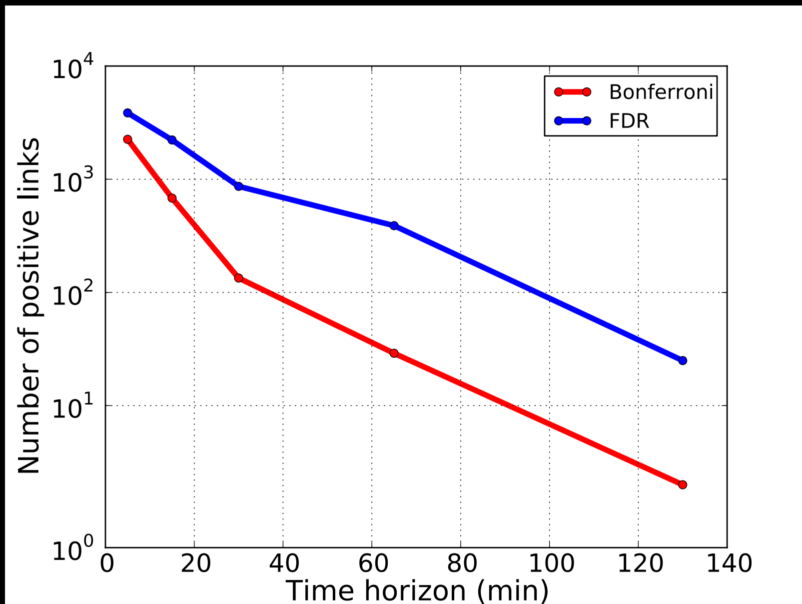
Network formed using returns sampled at Δt = 30 min.
Transaction data are from 2011-2012.



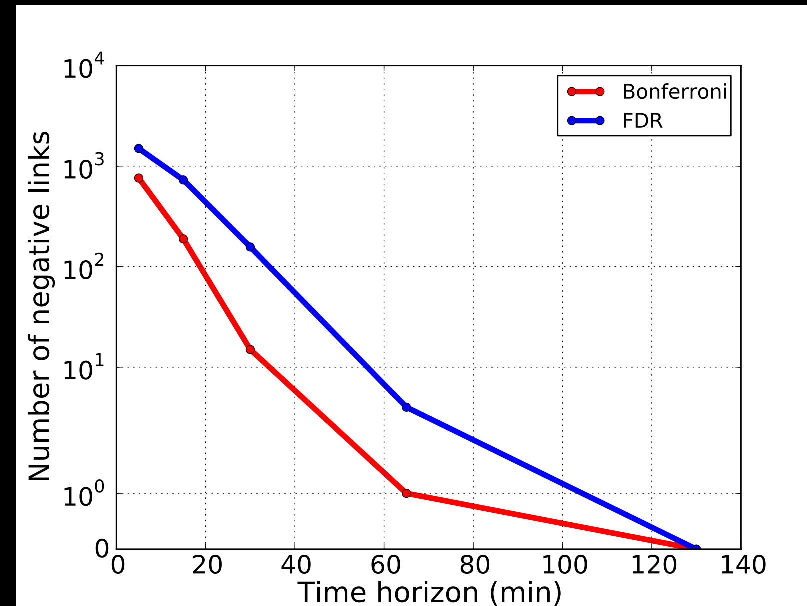
Network formed using returns sampled at Δt = 15 min.
Transaction data are from 2011-2012.



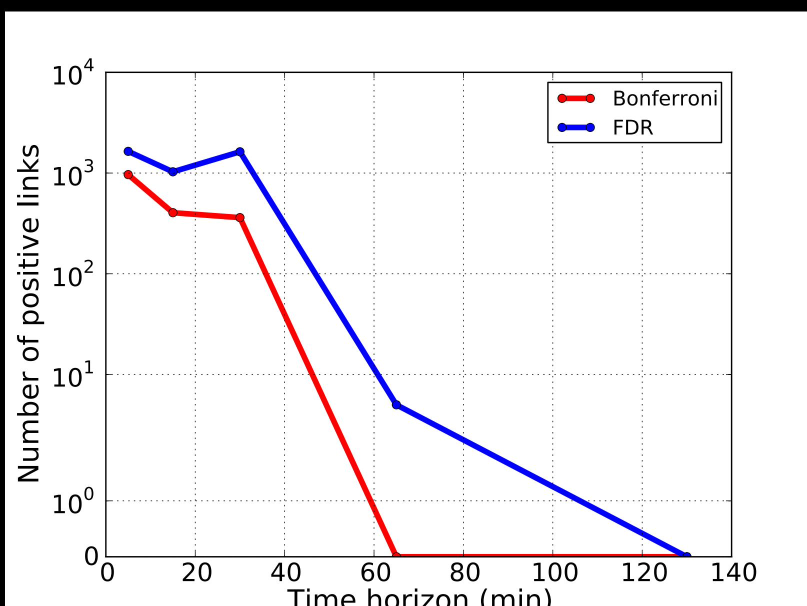
Network formed using returns sampled at Δt = 5 min.
Transaction data are from 2011-2012.



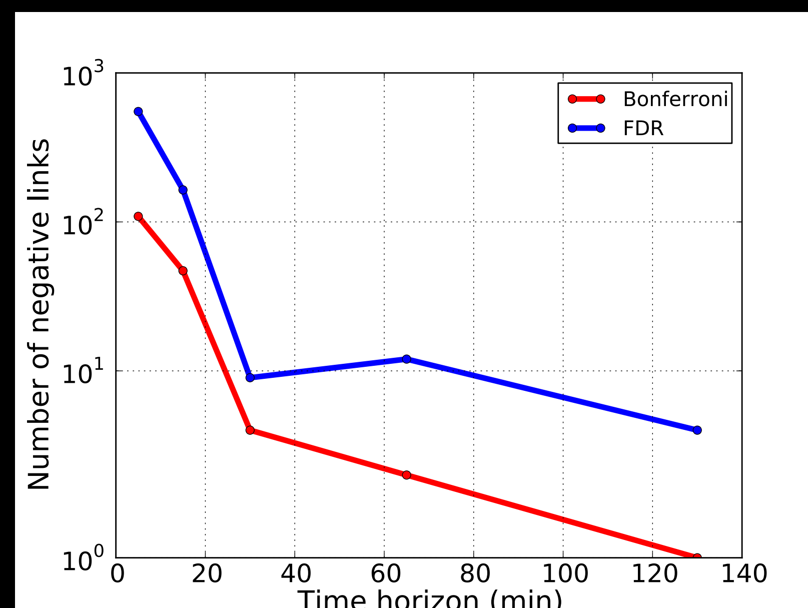
Positive links, 2002-2003



Negative links, 2002-2003

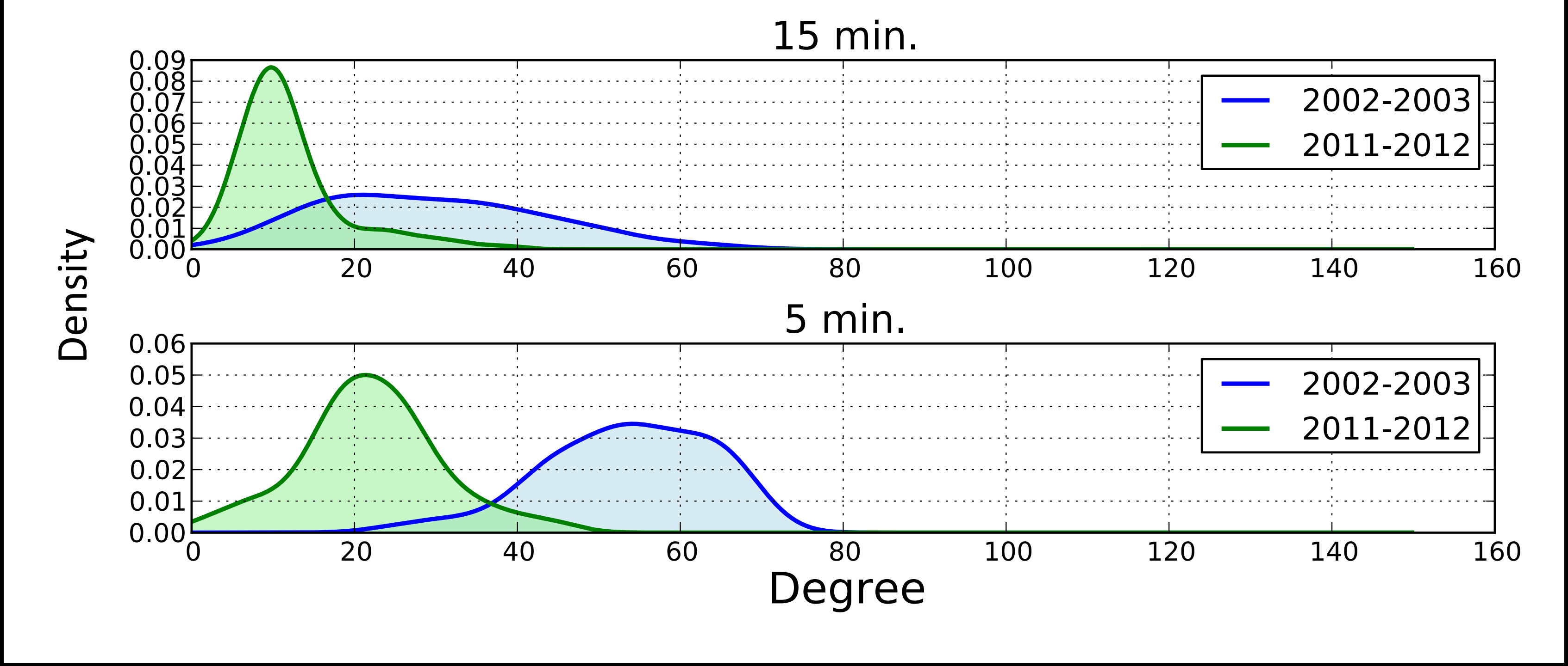


Positive links, 2011-2012



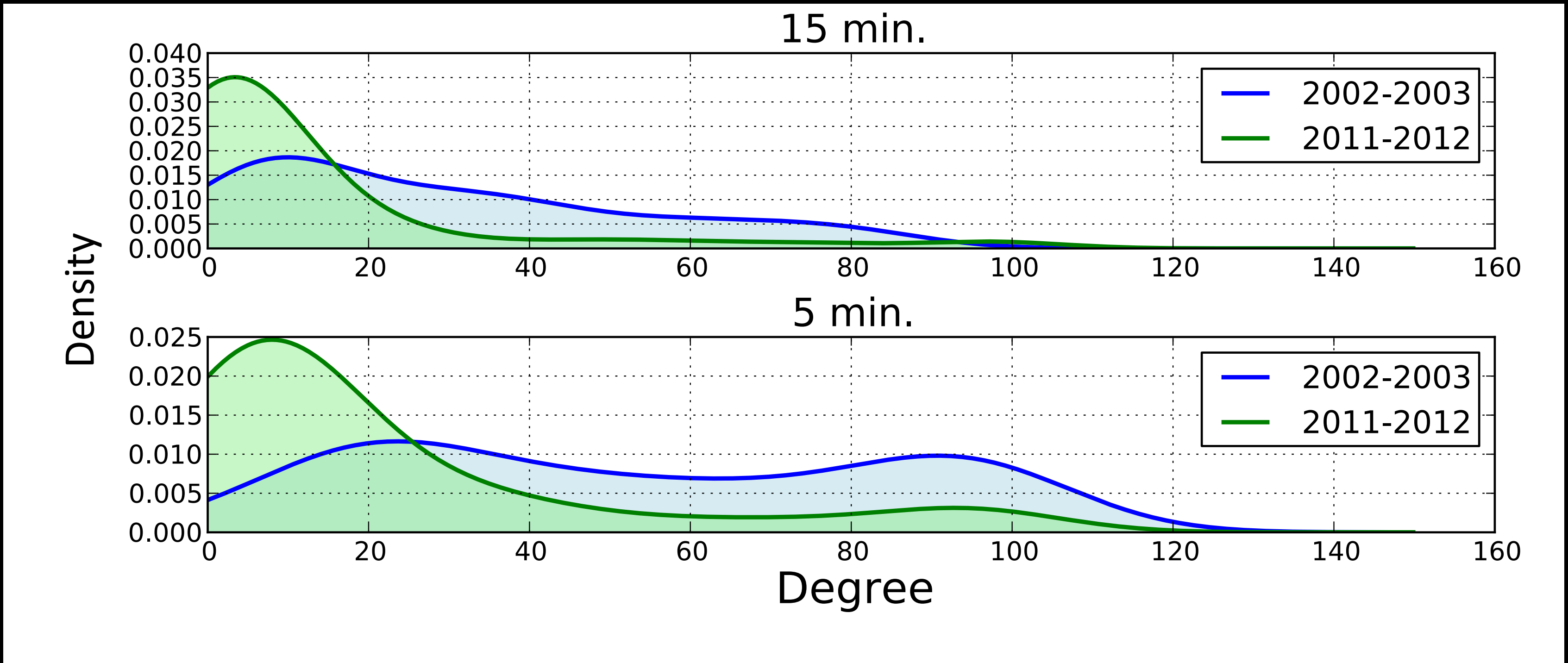
Negative links, 2011-2012

Market efficiency: leaders



Out-degree distributions (FDR networks)

Market efficiency: followers



In-degree distributions (FDR networks)

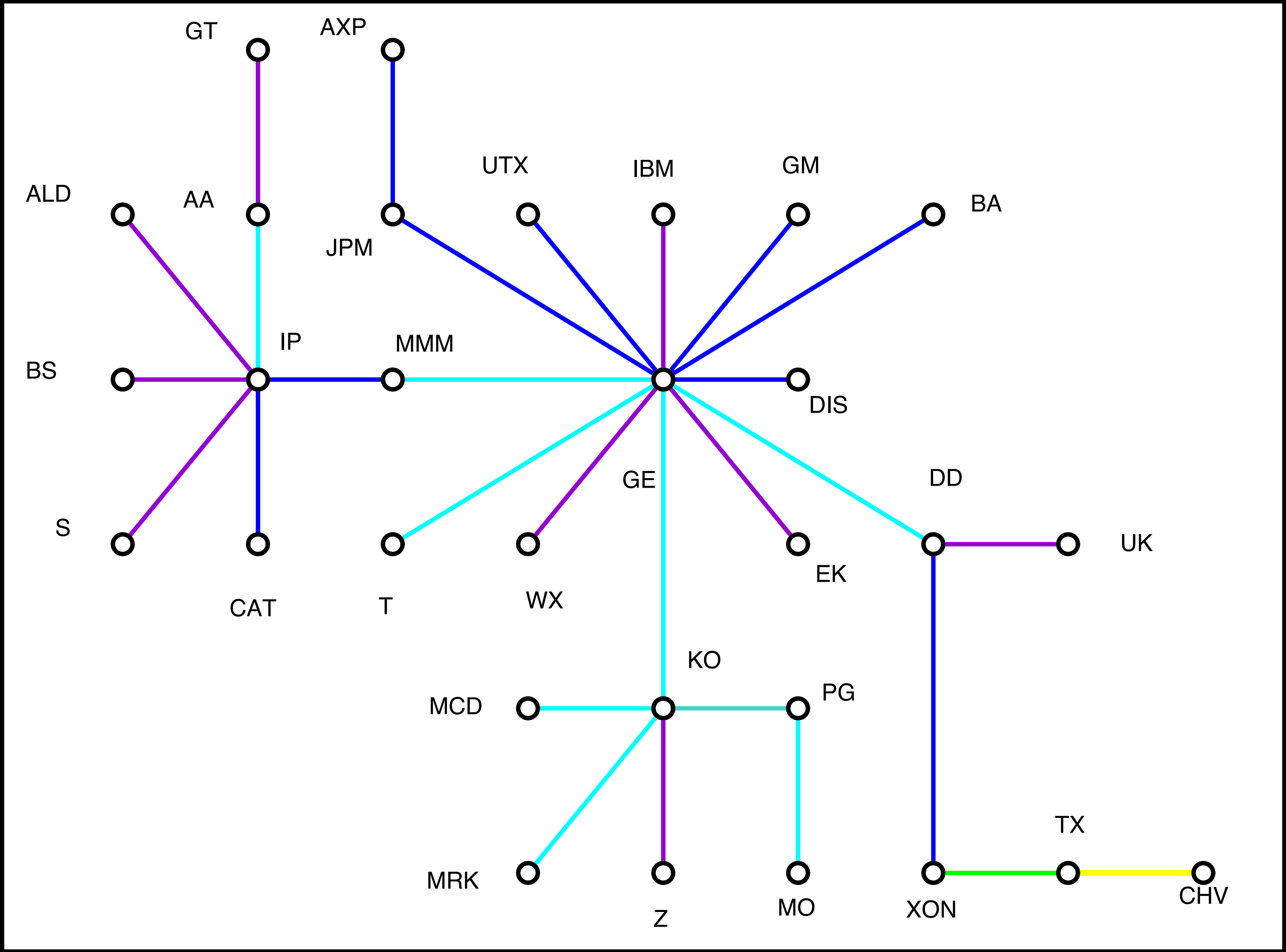
Comparison with threshold method

Δt	# pos. valid. (bootstrap)	# pos. valid. (normal dist.)	# pos. valid (both)
5 min	2,252	2,398	2,230
15 min	681	754	666
30 min	134	158	131
65 min	29	43	26
130 min	2	3	2

Ongoing work

- How might lagged correlations contribute to phenomena observed using synchronous correlations?
  - Economic sector clustering
- What changes occur during the trading day?

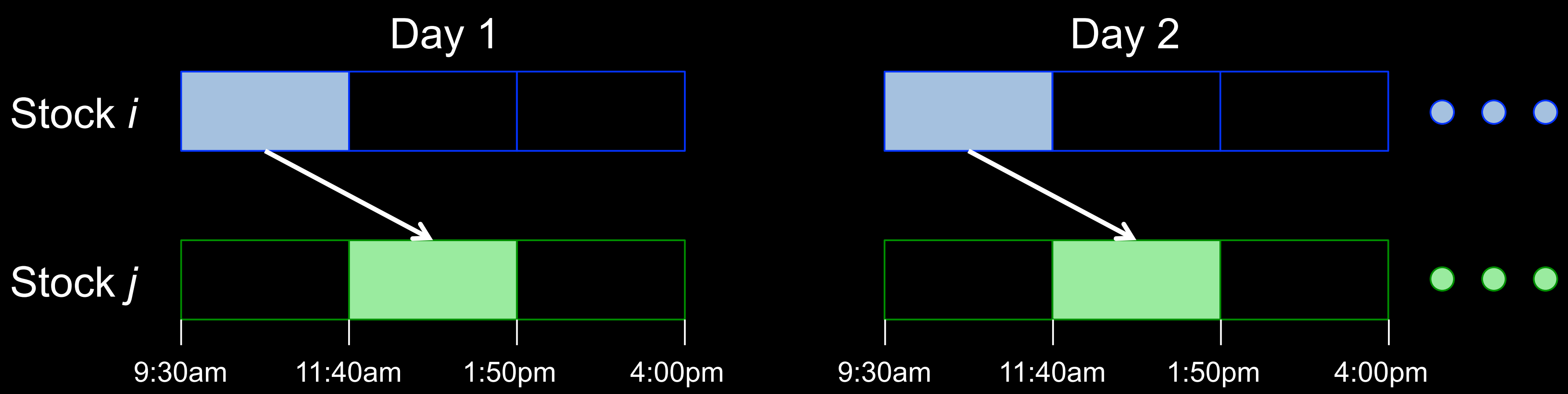
Hierarchical Method: Minimal Spanning Tree



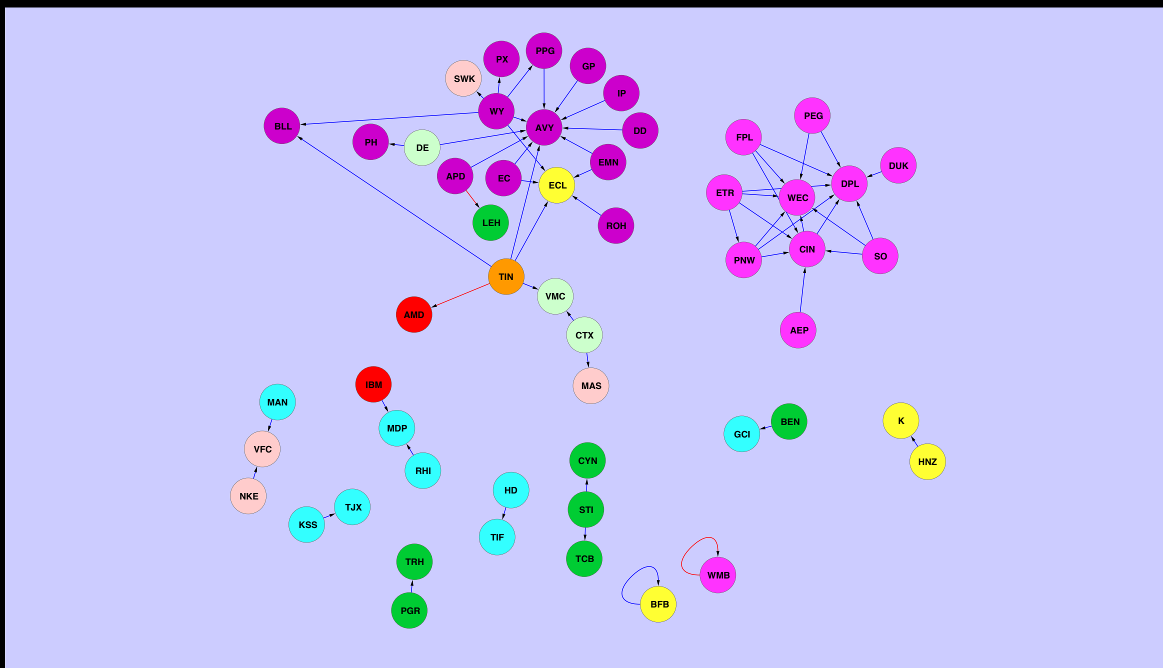
Mantegna 1999

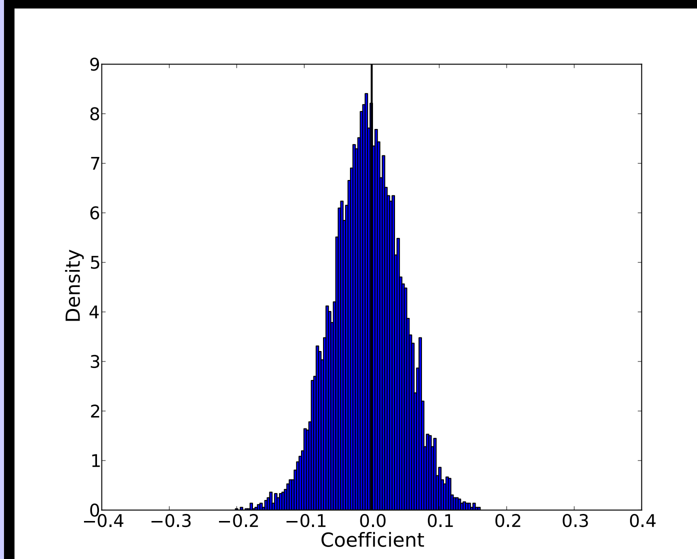
Intra-day seasonalities

Δt = 130 min:



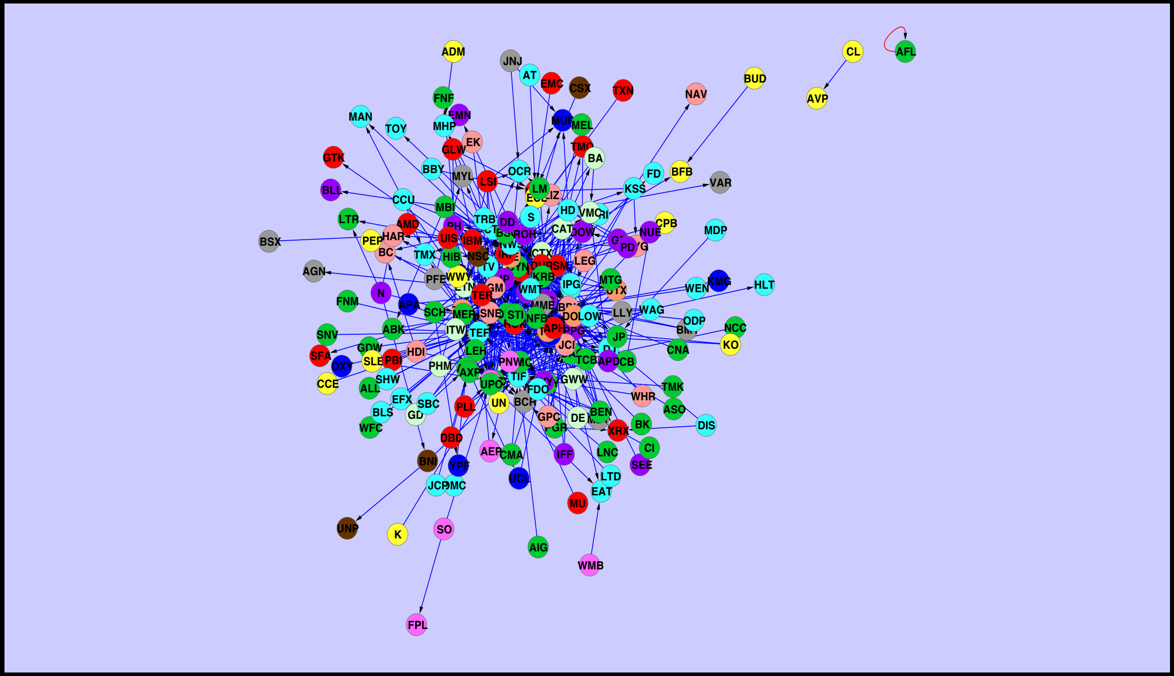
What changes occur in the course of a trading day?

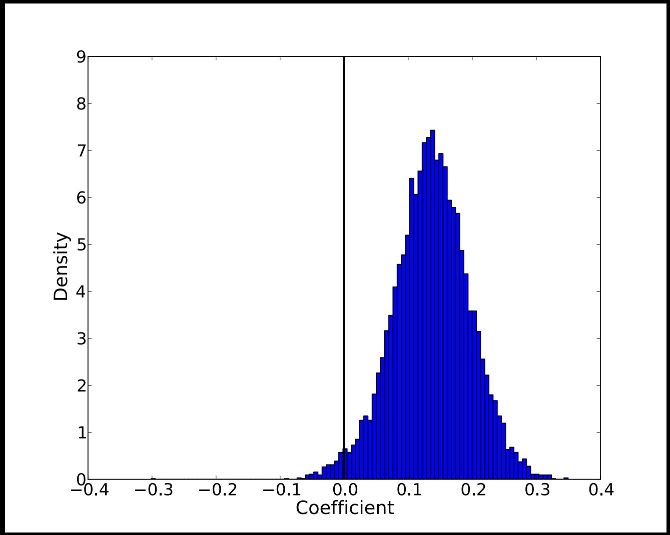




Correlations between returns in first 15 min. of trading day with returns in second 15min. of trading day.

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 Δt.
  - Increase in market efficiency over the past decade.

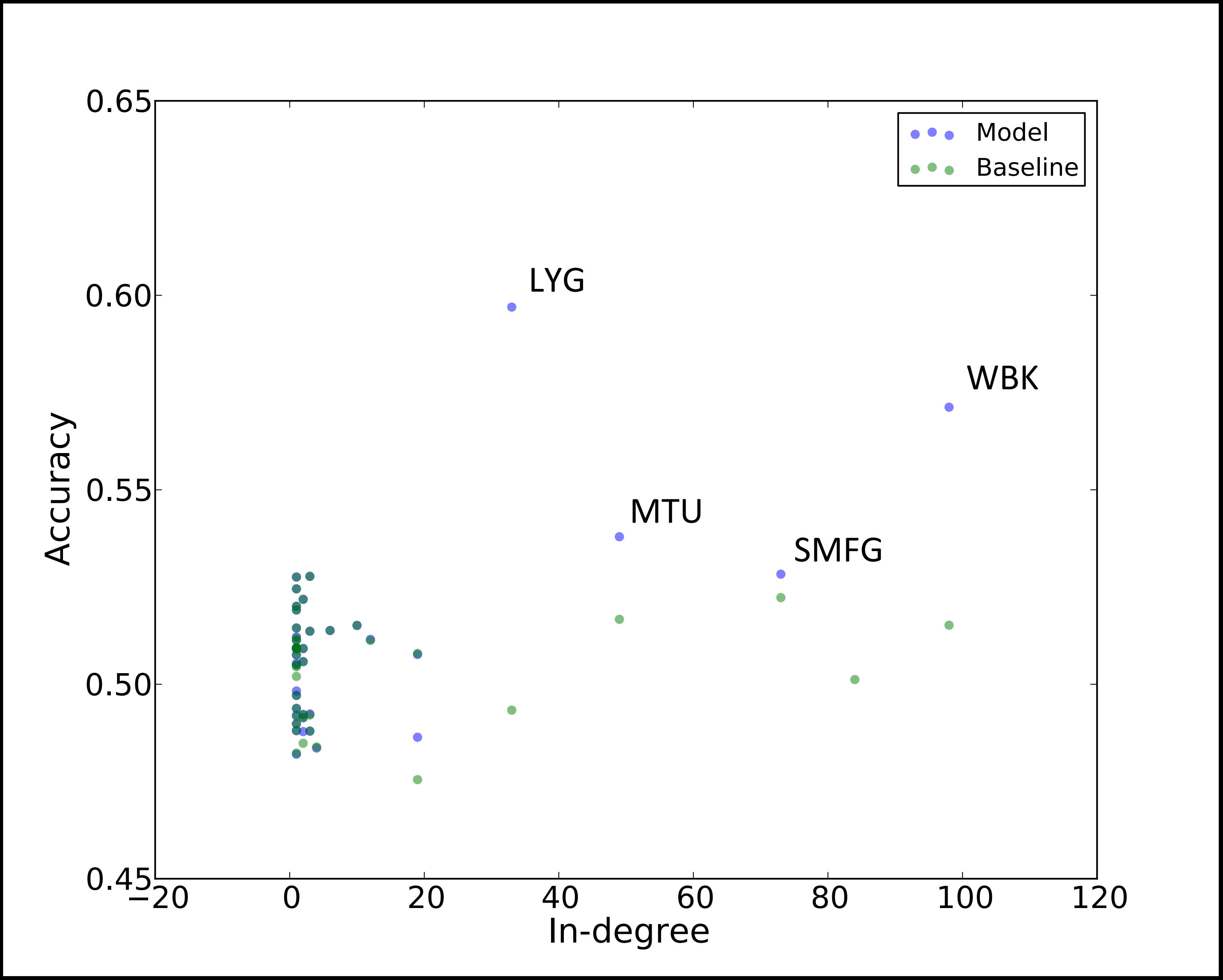
Future work

- 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!

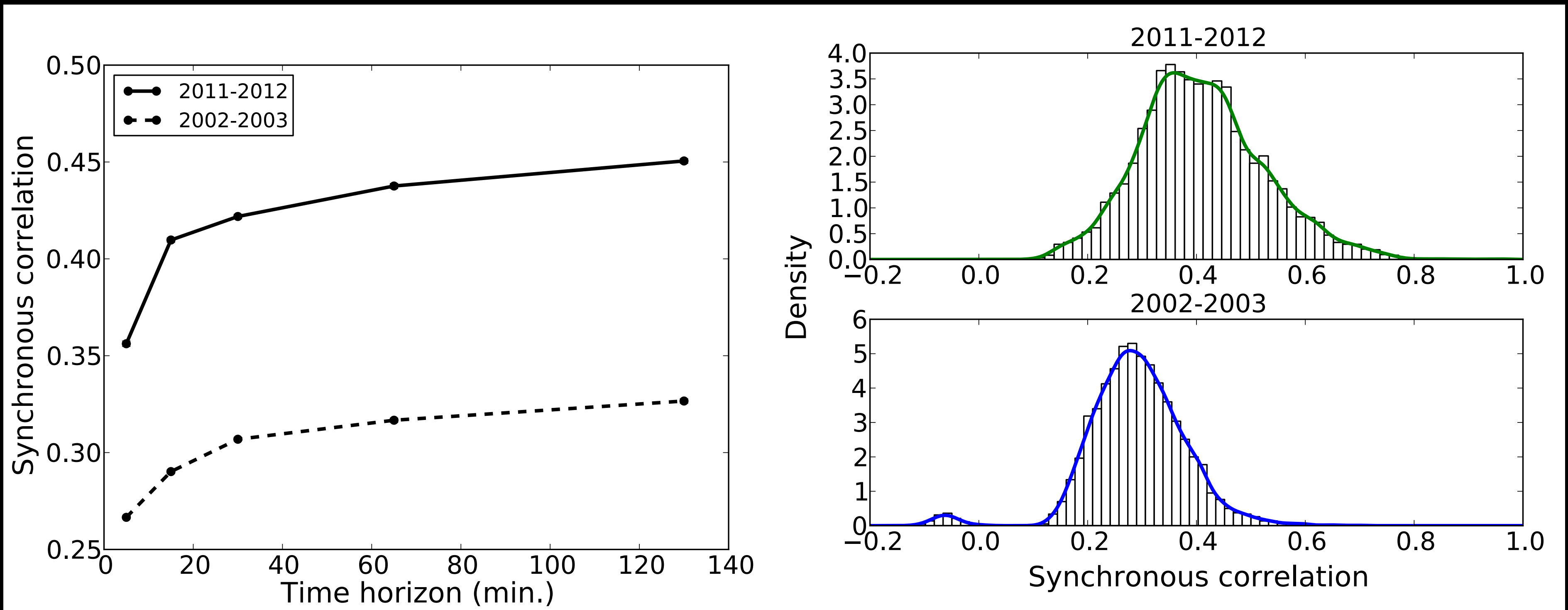
References

- Benjamini, Y. and Hochberg, Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1995, pp. 289-300.
- Bonanno, G., Lillo, F. and Mantegna, R.N., High-frequency cross-correlation in a set of stocks. *Quantitative Finance*, 2001, 1, 96-104.
- Efron, B. and Tibshirani, R., *An introduction to the bootstrap*, Vol. 57, 1993, CRC press.
- Epps, T., Comovements in stock prices in the very short run. *Journal of the American Statistical Association*, 1979, pp. 291-298.
- Gopikrishnan, P., Plerou, V., Liu, Y., Amaral, L., Gabaix, X. and Stanley, H.E., Scaling and correlation in financial time series. *Physica A: Statistical Mechanics and its Applications*, 2000, 287, 362-373.
- Havlin, S., Kenett, D.Y., Ben-Jacob, E., Bunde, A., Cohen, R., Hermann, H., Kantelhardt, J., Kertesz, J., Kirkpatrick, S., Kurths, J. et al., Challenges in network science: Applications to infrastructures, climate, social systems and economics. *European Physical Journal-Special Topics*, 2012, 214, 273.
- Hayashi, T. and Yoshida, N., On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 2005, 11, 359-379.
- Kenett, D.Y., Tumminello, M., Madi, A., Gur-Gershgoren, G., Mantegna, R. and Ben-Jacob, E. Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. *PLoS One*, 2010, 5, e15032.
- Malkiel, B.G. and Fama, E.F., Ecient Capital Markets: A Review Of Theory And Empirical Work. *The journal of Finance*, 1970, 25, 383-417.
- Mantegna, R., Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 1999, 11, 193-197.
- Song, D., Tumminello, M., Zhou, W. and Mantegna, R., Evolution of worldwide stock markets, correlation structure, and correlation-based graphs. *Physical Review E*, 2011, 84, 026108.
- Tumminello, M., Aste, T., Di Matteo, T. and Mantegna, R., A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America*, 2005, 102, 10421.
- Tumminello, M., Di Matteo, T., Aste, T. and Mantegna, R., Correlation based networks of equity returns sampled at different time horizons. *The European Physical Journal B-Condensed Matter and Complex Systems*, 2007b, 55, 209-217.



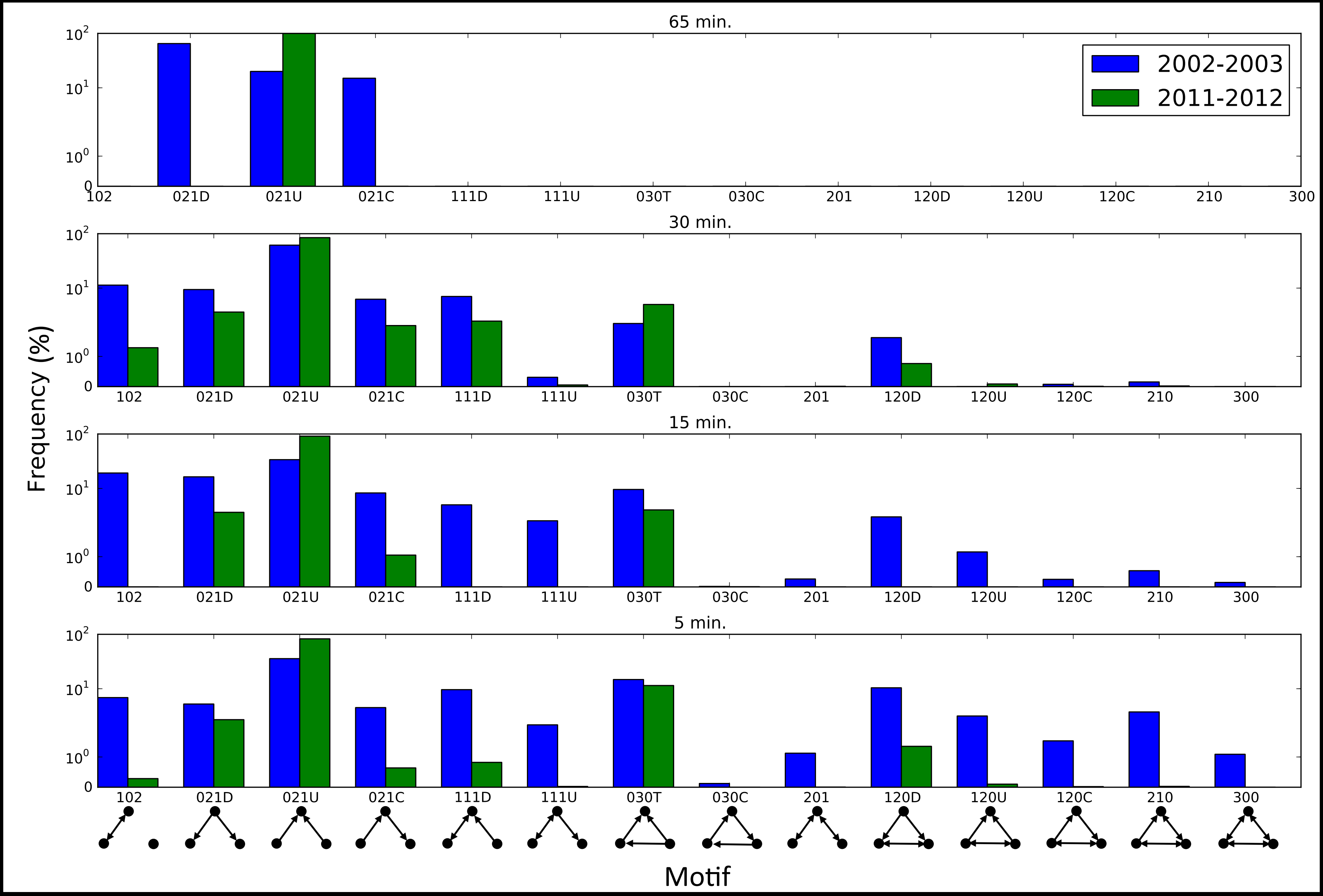
Accuracy of a logistic regression trained using 2011-2012 networks and tested on data from 2013.

Epps effect



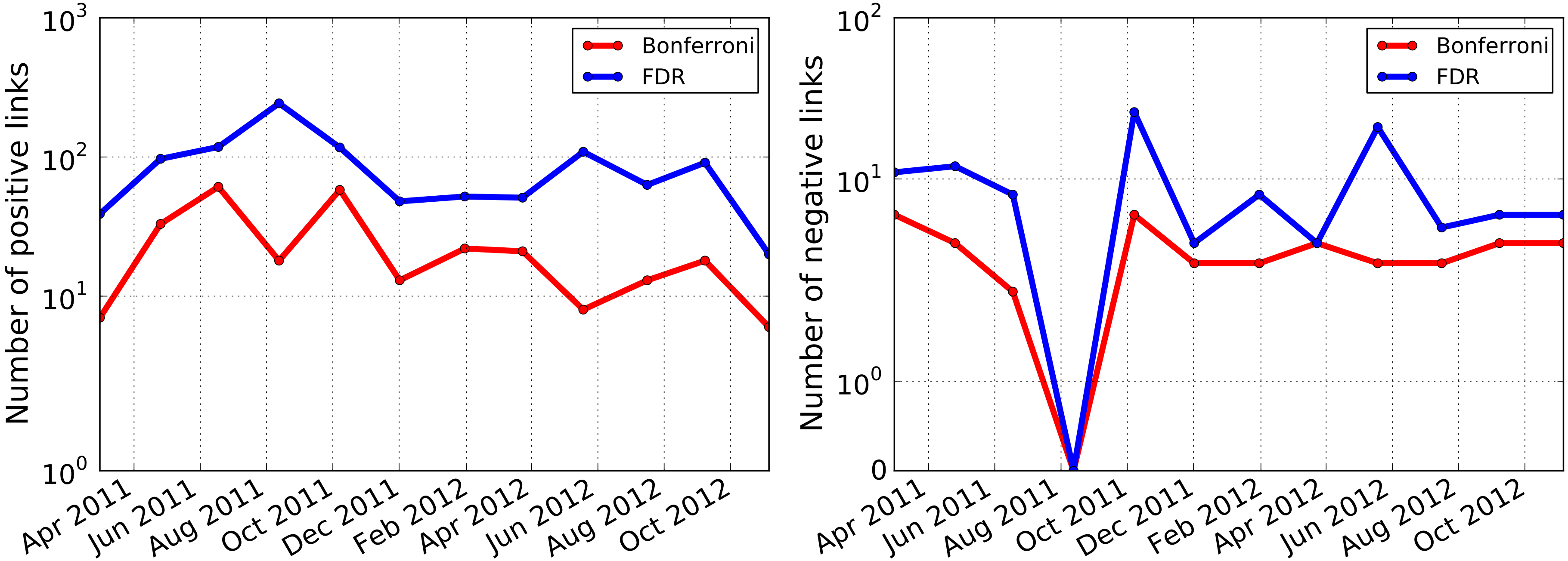
Epps curves of all synchronous correlation coefficients.

Distributions of correlations at Δt = 15 min.

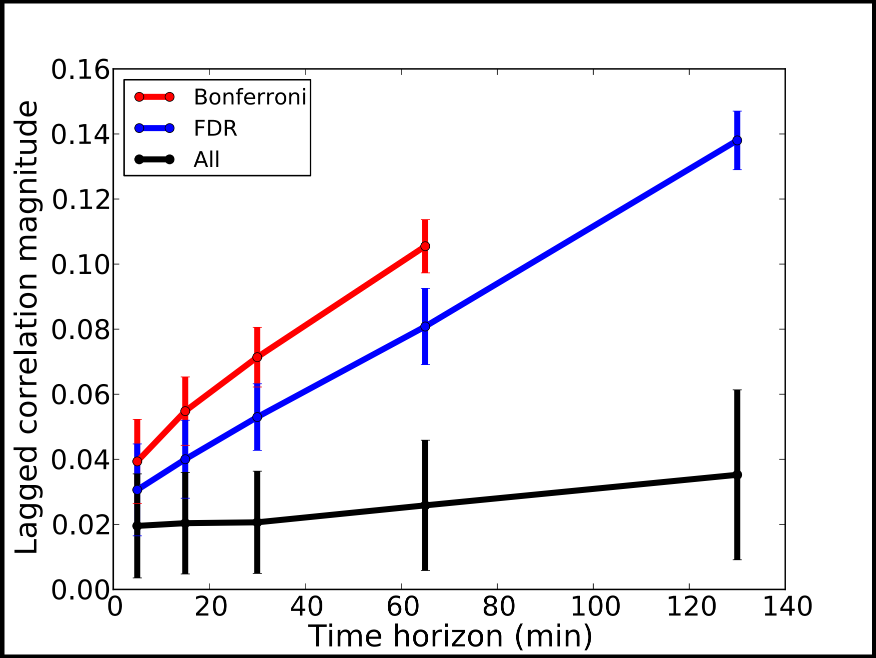


Motif counts in Bonferroni networks (2011-2012).

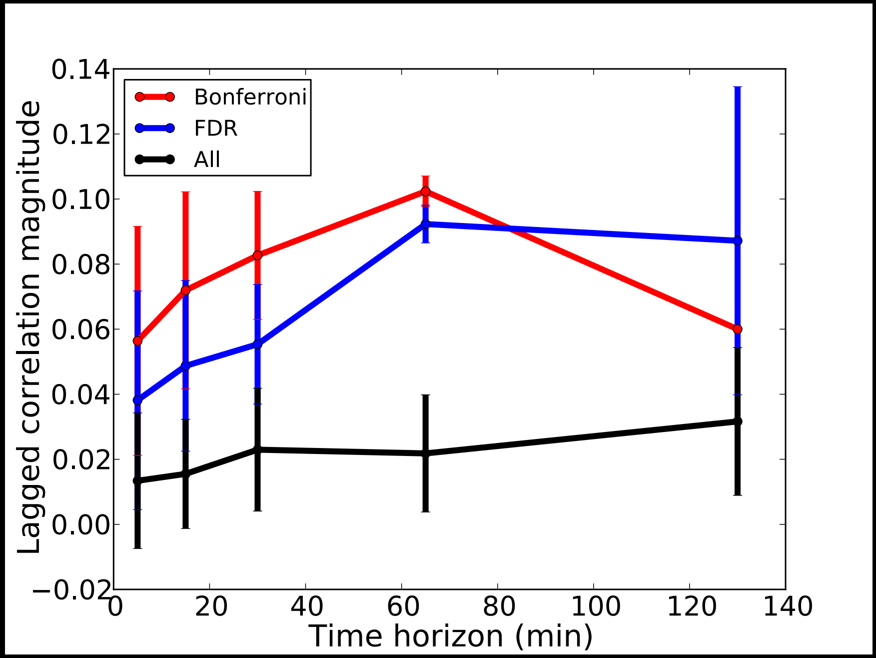
Effect of time series length on statistical power at a fixed time horizon Δt = 15 min.



Magnitude of filtered correlations

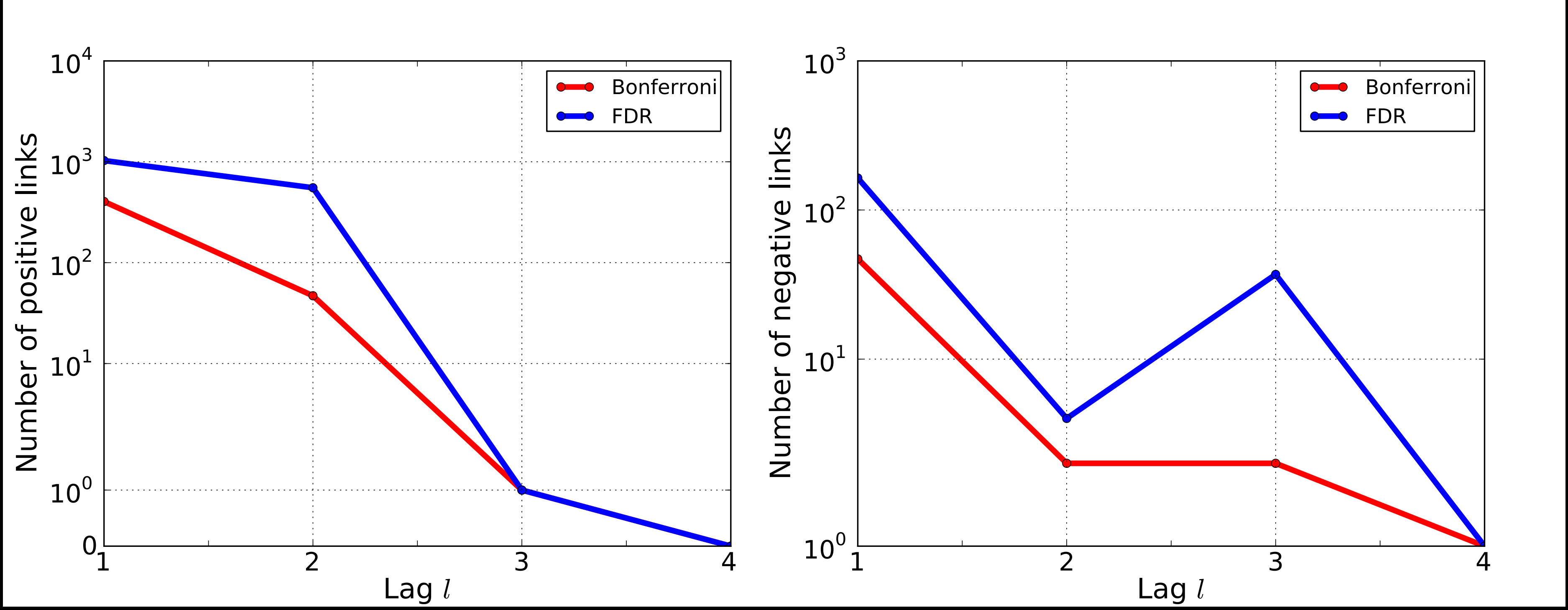


2002-2003

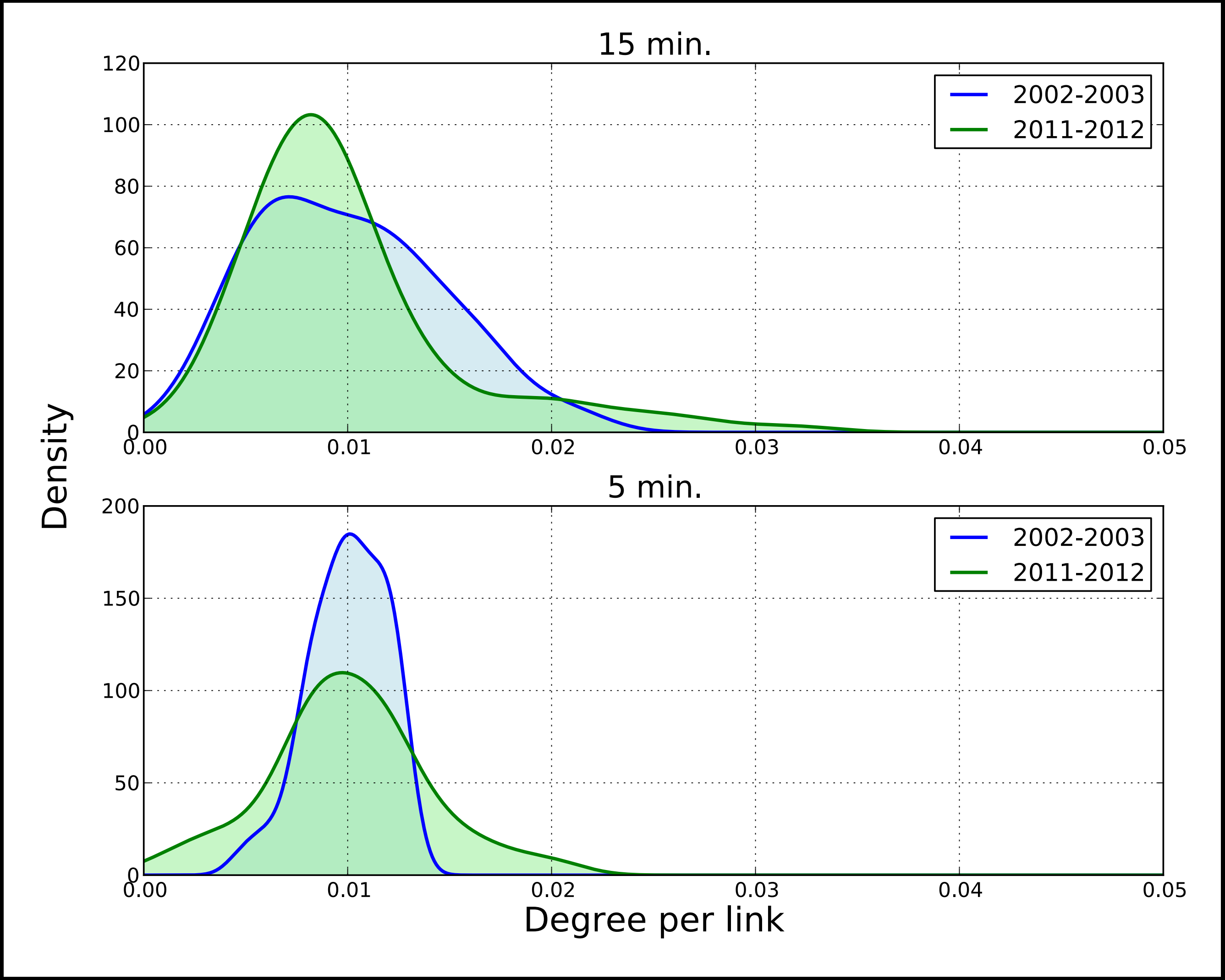


2011-2012

Effect of changing time lag at a fixed time horizon Δt = 15 min.

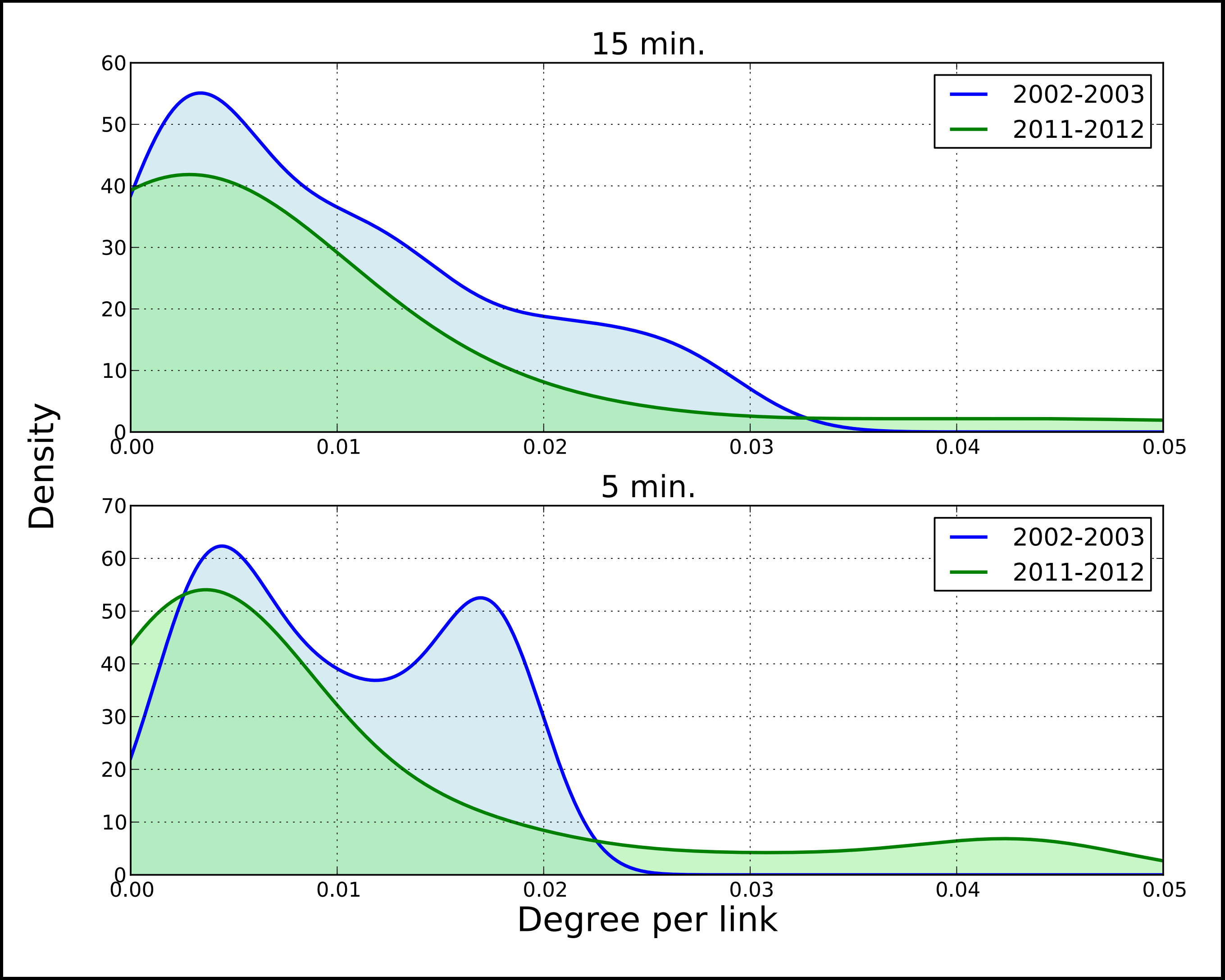


Market efficiency



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

Market efficiency



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