Correlation structure of spiky financial data: the case of congestion in Day-Ahead Energy Markets

F. Caravelli[1, 2, 3]

[1]*Invenia Technical Computing, 135 Innovation Dr., Winnipeg, MB R3T 6A8, Canada*
[2]*Department of Computer Science, UCL,Gower Street, London, WC1 E6BT, UK*
[3]*London Institute of Mathematical Sciences, 35a South Street, London W1K 2XF, UK*

I study the correlation structure and argue that these should be filtered. I propose the use of different correlation measures other than Pearson, in particular a modification of Event Synchronization adapted to negative values or a filtered correlation matrix.

Invenia Technical Report No. 2.

I. INTRODUCTION

Energy markets are notorious for being hardly predictable and volatile and with a bidding procedure which is not very transparent. In this paper we will study the correlation structure for two PJM and MISO using various methods, and we will use different methods to uncover some underlying structure.

II. DATA STRUCTURE

The data I study in the present paper has been collected by Invenia Technical Computing Co. and used in [1]. This data is available for each node in the PJM and MISO markets for a period of time of 1632 hours (January to March), starting the 1st of January 2014 00:00am; in particular, there are 1287 nodes for PJM and 2568 nodes in MISO. For each node in the DA markets there are 3 time series, which is an aggregated and averaged price of the previous hour, from a 5-minutes interval forecast in day-ahead. The full price of electricity at node n, at time t is given by:

$$LMP(n,t) = MEC(t) + MCC(n,t) + MLC(n,t), \quad (1)$$

where MEC stands for *Marginal Energy Component*, MCC stands for *Marginal Congestion Component* and MLC stands for *Marginal Loss Component*. The $MCC(n,t)$ component is the price due to transmission congestion, i.e., it is the marginal cost of supplying the next increment of load at a location, taking into account the transmission constraints of the grid. This can be positive and negative, and is often 0, and can be thought as the equivalent of returns in stock markets. The MCC component of the price at a node takes into account the costs incurred due to the physical constraints of the transmission system. For example, when a power line at some location is at its limit for carrying power, the load at that location must be serviced through another line, which can be more costly. $MLC(n,t)$ is the price due transmission losses on the grid. This is generally small as compared to MEC. $MEC(t)$ is the price of electricity at any given node if there is no congestion and loss to that node.In general, the MEC component of the time series is independent of the node, and thus represents a price shift for the whole market. The MLC and MCC components are instead time and node dependent, but we observe that in general $MCC \gg MLC$. Thus, for the present paper we study only the MCC component of the prices in the day ahead market, which directly address the inefficiency of power transmission and is the main source of volatility in the LMP time series.

III. CORRELATION AND SYNCHRONIZATION MEASURES

The correlation structure for the markets can be obtained through standard Pearson correlation. In general, we will be interested in the structure arising from correlations both in the case of Pearson and in the case of a modified version of Event Synchronization.

A. Correlation

Empirical cross-correlation is the most used device to estimate interdependence between time series. These assume stationarity and linear interdependence. The correlation matrix is defined as:

$$C_{ij} = Corr[X_i, X_j]\sqrt{Var[X_i]Var[X_j]} \quad (2)$$

with

$$Cov[X_i, X_j] = \langle X_i X_j\rangle - \langle X_i\rangle\langle X_j\rangle \quad (3)$$

and $Var[X_i] = \sigma_i^2 = \langle X_i^2\rangle - \langle X_i\rangle^2$. Averages are taken as $\langle X_i\rangle = T^{-1}\sum_{t=1}^{T} x_i(t)$.

Pearson correlation relies both on the stationarity of the time series. Pearson correlation is shown in Fig. 1 (top) for MISO and in Fig. 2 (top) for PJM.

B. Synchronization

Standard event synchronization. Given the spiky nature of the data to be analyzed, we consider another dis-

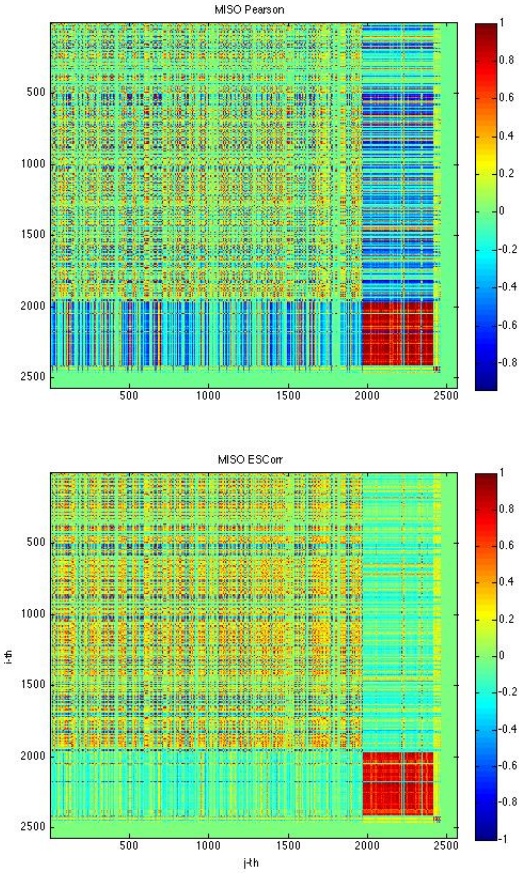

FIG. 1: Pearson correlation (top) and Synchronization (bottom) for MISO.

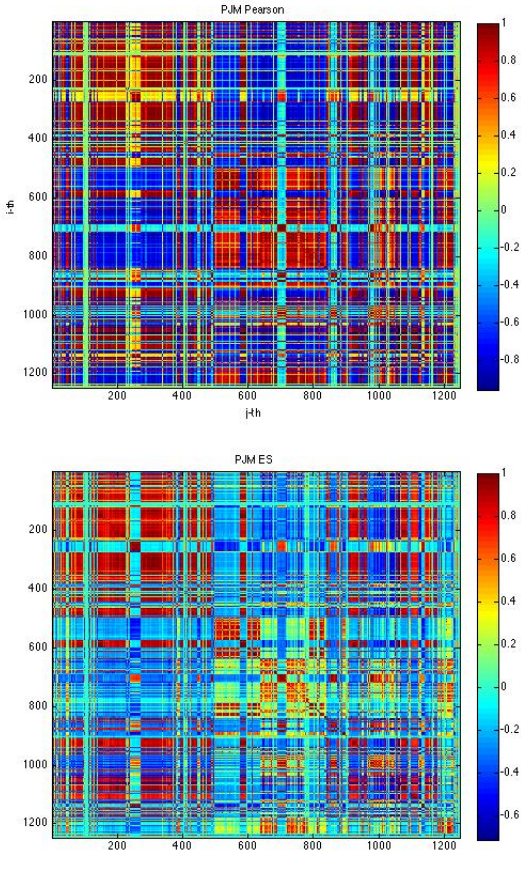

FIG. 2: Pearson correlation (top) and Synchronization (bottom) for PJM.

tance between time series which is a variation of event synchronization. We denote with $c^\tau(x|y)$ the number of times an event appears in x shortly after it appears in y,

$$c^\tau(x|y) = \sum_i \sum_j J_{ij}^\tau \tag{4}$$

where $J_{ij} = 1$ if $0 < |t_i^x - t_i^y| \leq \tau$, $J_{ij} = 1/2$ if $t_i^x = t_i^y$ and zero otherwise. Where one writes $Q_\tau = \frac{c^\tau(y|x)+c^\tau(x|y)}{\sqrt{m_x m_y}}$. Event synchronization has been constructed to measure the amount of synchronization between two (spiky) time series, in particular for neural networks.Q_τ is not a measure of correlation or anti-correlation.

Given the spiky behavior of time series of marginal congestion, one could ask himself is such a measure of synchronization; this can be done as follows.

Inspired by this, we consider a variation to this approach to account for positive and negative spikes. First of all, we consider a filtering process for fat tailed data set. Given a time series x made of positive and negative data, we introduce the positive and negative thresholds mp_x and mn_x as the median of the positive values and the median of the negative values of x respectively, calculated in a time window of length T. We then consider a positive "event" one which is above the threshold mp_x, and a negative ones those below mn_x. Using these thresholds, we convert the time to a series of 0, -1 and 1's; we define this transformed time series as $\epsilon_x(t)$. This filtering is depicted in Fig. III B. At this point, J_{ij}^τ is constructed as follows:

$$J_{ij}^\tau = \sum_t \epsilon_{x_i}(t) \sum_{t',|t-t'|\leq\tau} \epsilon_{x_j}(t') \tag{5}$$

we then consider the matrix $D_{ij} = \delta_{ij} J_{ij}$, and normalize the synchronization matrix J_{ij}, obtaining an alternative definition of correlation matrix:

$$C' = \sqrt{D^{-1}} J \sqrt{D^{-1}} \tag{6}$$

which is symmetric by construction. This measure is shown in Fig. 1 (bottom) for MISO and in Fig. 2 (bottom) for PJM for $\tau = 3$ h.

C. Differences between the two methods

As a first comment, we note that the two correlation matrices build a correlation matrix which are similar, but

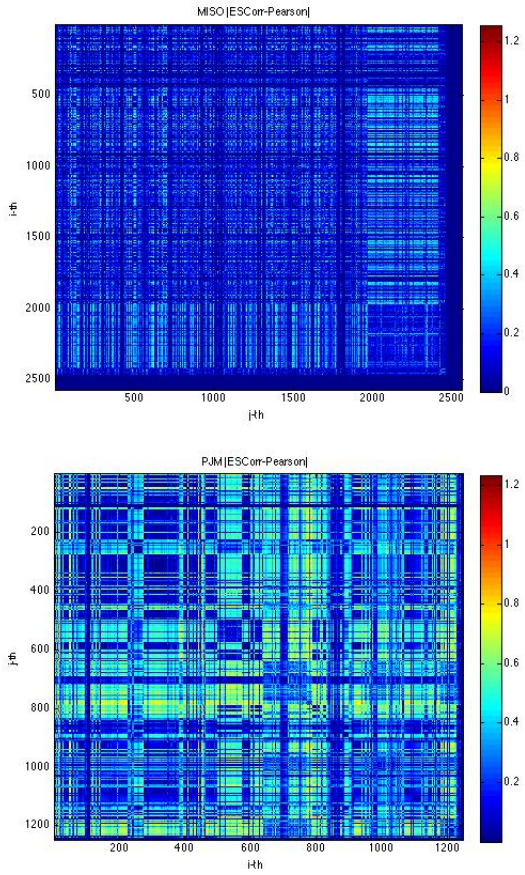

FIG. 3: Difference between the two correlation matrices methods for MISO (top) and PJM (bottom).

different. In Fig. 3 we note how the two methods obtain different results for MISO (top) and PJM (bottom), and that this different is in particular visibly stronger in the case of PJM.

IV. FILTERING METHODS

A. Network-based clustering algorithms

A first approach to detect structure in Markets is the use of Clustering algorithms. As a first example, I consider the Girvan-Newman algorithm [2], which detects communities by progressively removing edges from the original network; then the connected components of the remaining network are considered to be the eventual communities. The Girvan-Newman algorithm focuses on edges that are most likely "between" communities. An application of this algorithm to the case of Pearson correlation when correlations lower than 0.7 have been neglected is shown in Fig. 10 for PJM (a) and for MISO (b).

B. MFT and PMFG

Another filtering approach looks for the Minimal Spanning Tree (MST) obtained again from the strongest correlations, but now retaining only the N -1 correlations that are required for each node to be reachable from any other node via a connected path, while discarding those that produce loops. This procedure automatically produces an agglomerative hierarchical clustering (a dendrogram) of the original time series and requires that the correlation matrix is renormalized at each iteration of the clustering according to some protocol, until a final filtered matrix is obtained [3–5].The MST method does not require the introduction of an arbitrary threshold, but it assumes that the original correlations are well approximated by the filtered ones. At a geometrical level, this corresponds to the assumption that the metric space in which the original time series are embedded (via the definition of a proper correlation-based distance) effectively reduces to a so-called ultrametric space where well-separated clusters of points are hierarchically nested within larger well separated clusters. Even if the method exploits the correlations required for the MST to span the entire set of time series, it discards all the weaker correlations. Moreover, the approximating correlations are progressively more distant from the original ones as higher and higher levels of the taxonomic tree are resolved. This means that the method is more reliable when using the strongest correlations to determine the low-level structure of the taxonomic tree (small clusters of time series), while it is progressively less reliable when using the weaker correlations to determine a taxonomy. An alternative approach, which is similar in spirit to the MST but discards less information, is the so-called Planar Maximally Filtered Graph (PMFG) [6]. This method allows one to retain not just the correlations required to form the MST, but also a number of additional ones, provided that the resulting structure is a planar graph (a network that can be drawn on a plane without creating intersecting links). A nice feature of the PMFG is that it always contains the entire MST, so that the former provides additional, and not just different, information with respect to the latter. However, also this method is affected by some degree of arbitrariness, which lies again in the properties of the postulated, approximating structure. There is no obvious reason why stocks (or other time series) should find a natural embedding in a bidimensional plane. In fact, the PMFG has also been described as the simplest case of a more general procedure based on the embedding of high-dimensional data in lower-dimensional manifolds with a controllable genus (number of "handles" or "holes"). The PMFG corresponds to the case when the genus is zero. So the arbitrariness of the method can be rephrased as its dependence on some value of the genus that must be fixed a priori. The method has been extended in a variety of ways in order to produce a nested hierarchy of time series by exploiting the properties of the embedding space.

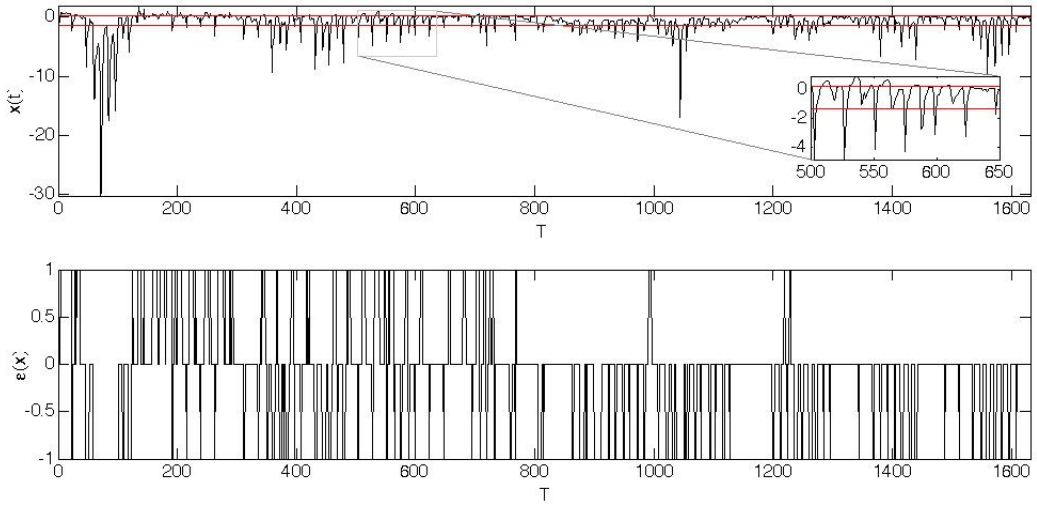

FIG. 4: Filtering the time series according to its positive and negative median value.

However, as with the MST, the target of these methods is that of finding the postulated approximating structure, rather than optimizing the search of groups of time series that are more correlated internally than with each other.

Such procedure has been performed for MISO and PJM in Fig. 9.

C. Random matrix theory

We finally mention an important technique, based on random matrix theory (RMT), which is widely used in order to identify the non-random properties of empirical correlation matrices [7]. We will use this technique extensively in the paper. A correlation matrix constructed from N completely random time series of duration T has (in the limits $N \to \infty$ and $T \to \infty$ with $1 < \frac{T}{N} < \infty$) a very specific distribution of its eigenvalues, known as the Marcenko-Pastur or Sengupta-Mitra distribution. This distribution reads

$$\rho(\lambda) = \frac{1}{\rho} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda_- - \lambda)}}{2\pi\lambda} \tag{7}$$

and

$$\lambda_\pm = (1 + \sqrt{\rho})^2 \tag{8}$$

and $\rho = \frac{T}{N}$. In general, one assumes that the correlation matrix can be decomposed into two parts, $C = C^r + C^s$, where C^r is the 'random' component and C^s is the matrix containing information.

The deviation of the spectra of real correlation matrices from the RMT prediction provides an effective way to filter out noise from empirical data, and also illustrates some robust property of financial markets. For instance, in insets of Fig. 5 we superimpose the eigenvalue density of the empirical correlation matrix obtained from MISO and PJM and the corresponding expectation given by the Marcenko-Pastur distribution with the same values of N and T. It has been shown in different studies that a typical feature of the spectrum of empirical correlation matrices is that the largest observed eigenvalue M is much larger than all other eigenvalues, as in Fig. 5 for the case of PJM and MISO. Rapid changes in the eigenvector implies that the correlation structure is changing sharply and that some transition phenomena are occurring in the market. The corresponding eigenvector has often all positive signs and one can therefore identify this eigencomponent of the correlations as the so-called market mode, i.e. a common factor influencing all stocks within a given market. Interpreting this, the bulk of the correlation between pairs of stocks is attributed to a single common factor, much as all boats in a harbor will rise and fall with the tide. In order to clearly see which "boats" are rising and falling relative to one another, one must subtract out the common "tide", which in terms of the correlation matrix leads to the further decomposition

$$C = C^r + C^g + C^m; \tag{9}$$

this analysis is performed in Fig. 6. As a main finding, we observe that The correlations embodied by C^g act neither at the level of individual stocks (uncorrelated noise), nor at that of the entire market. Such correlations act at the level of sub-groups of stocks within a market, and they are often referred to as the "group" mode. The eigenvectors contributing to C^g have alternating signs, and this allows the identification of groups of stocks that are influenced in a similar manner by one or more common factor.

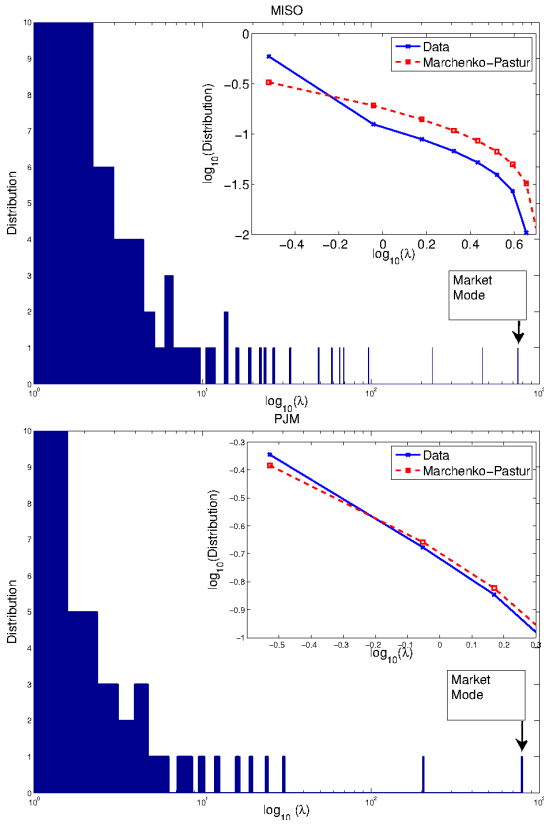

FIG. 5: Pearson Correlation matrix spectrum for MISO (top) and PJM (bottom). Inset graphs are comparisons with Marchenko-Pastur against those obtained after reshuffling.

D. Difference between event synchronization and Filtered RMT

One of the main outcome of the previous section is that correlation matrices very often have a content which indeed dominated by random fluctuations. These fluctuations should be removed in order to unveil the "real" correlation structure. The outcome of the filtering procedure is shown in Fig. **??**. In MISO we observe that most of the anticorrelation between the upper and lower block disappears, and that the lower block correlation is less pronounced. For PJM, we observe that most of the information is indeed filtered, leaving only few strong correlation blocks.

Given the preamble on the importance of filtering the correlation matrix, we ask ourselves whether Event Synchronization is capturing some "real" effects. Being Event Synchronization a nonlinear measure of correlation, this might be well possible. However, an extra parameter of this approach is the "synchronization parameter" τ. We thus evaluate the average difference E between the filtered correlation matrix in Fig. 3 and the correlation measured by Event Synchronization; this average difference is shown in Fig. 8, and shows that for the case of PJM one needs to set the parameter τ to 21h, meanwhile for MISO this should be set to $\tau = 9$.

V. CONCLUSIONS

We have discussed various methods for filtering the correlation matrix and an alternative measure of correlation, based on a modification of Event Synchronization. We have argued that most of the correlation observed is due to random fluctuations in the market and that these should be filtered; alternatively, one should use nonlinear correlations, such as the one tried in this article. In addition to this, we have explored various methods for unveiling the correlation structure for PJM and MISO, identifying various clusters and which neglect most of the information present in the correlation matrix. These methods are robust, as they use only the strongest correlation links.

Appendix A: Modified Event Synchronization code

```
% [J, Z]=ESCORRmed(P,tau)
% It calculates correlation based upon
% the median threshold, tau is the time window
% J is the correlation, Z the spikes used
% WARNING
% the first dimension in P is nodes and the
%second is time
function [J, Z]=ESCORRqmed(P,tau)
N=size(P,1);
L=size(P,2);

Z1=zeros(size(P,1),size(P,2));
Z2=zeros(size(P,1),size(P,2));
for i=1:N

 Z1(i,:)=P(i,:)>nanmedian(P(i,P(i,:)>=0));
```

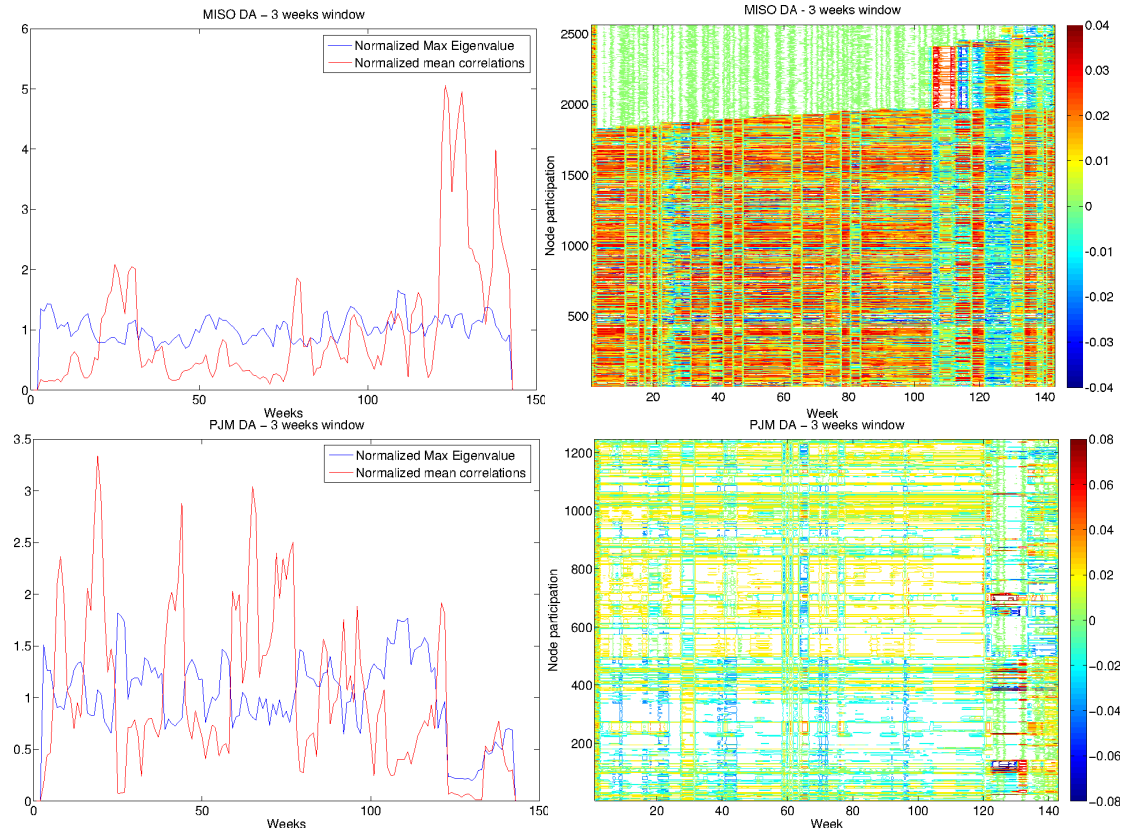

FIG. 6: Pearson Correlation matrix largest eigenvector as a function of time for MISO (top left) and PJM (bottom left), evaluated over a 3 weeks window for every week for the available data. We observe that for every peak in the eigenvector (blue) or the mean correlation (red), there is a sharp transition in the population of the main eigenvector (right).

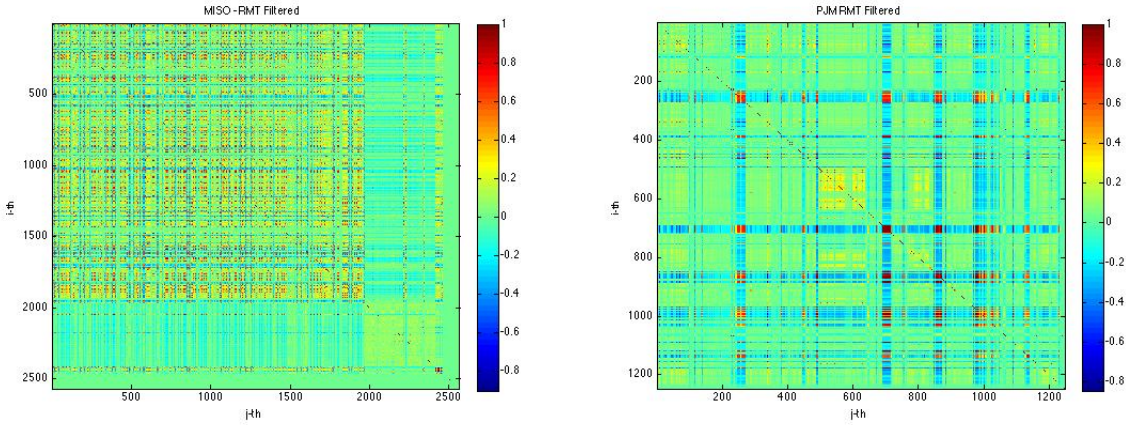

FIG. 7: Filtered Correlations evaluated using the RMT for MISO (left) and PJM (right). In MISO we observe that most of the anticorrelation between the upper and lower block disappears, and that the lower block correlation is less pronounced. For PJM, we observe that most of the information is indeed filtered, leaving only few strong correlation blocks.

```
 Z2(i,:)=P(i,:)<nanmedian(P(i,P(i,:)<=0));
end
Z=double(Z1)-double(Z2);
```

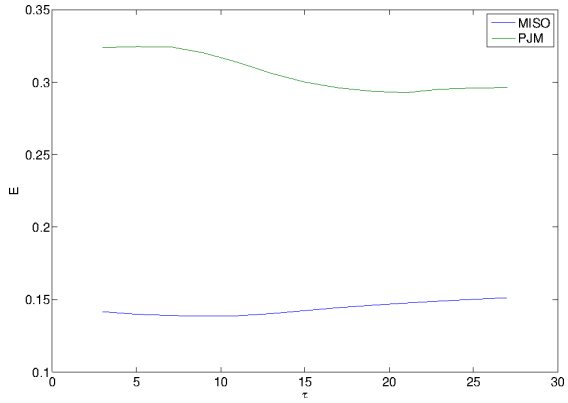

FIG. 8: Average difference E between the filtered RMT correlation matrix and event synchronization as a function of τ for MISO (blue line) and PJM (green line). Wee see that for both markets there is a minimum, obtained approximately at $\tau = 9$ for MISO and $\tau = 21$ for PJM.

```
clear Z1
clear Z2

D=zeros(N,L);
for k=1:L
    D(:,k)=sum(Z(:,max(k-tau,1):min(k+tau,L)),2);
end

J=D*D';
J=pinv(diag(sqrt(diag(J))))*J*pinv(diag(sqrt(diag(J))));


end
```

Appendix B: Filtered RMT

```
function [M, C] = FinRMT(diffs)
%% FinRMT
% FinRMT uses Random Matrix Theory (RMT) to create a filtered correlation
% matrix from a set of financial time series price data, for example the
% daily closing prices of the stocks in the S&P
%% Syntax
% M=FinRMT(priceTS)


N = size(diffs,2);      % N is the number of time series
T = size(diffs,1);      % T is the lenght of each series

diffs=RemoveNan(diffs')';


C = RemoveNan(corrcoef(diffs));    % Create a correlation matrix and ensure
C = .5 * (C+C');        % it's symmetric
```

```
[V,D] = eig(C);
[eigvals, ind]=sort(diag(D),'ascend');
V = V(:,ind);
D=diag(sort(diag(D),'ascend'));

Q=T/N;
sigma = 1 - max(eigvals)/N;
RMTmaxEig = sigma*(1 + (1.0/Q) + 2*sqrt(1/Q));
RMTmaxIndex = find(eigvals > RMTmaxEig);
if isempty(RMTmaxIndex)
    RMTmaxIndex = N;
else
    RMTmaxIndex = RMTmaxIndex(1);
end


RMTminEig = sigma*(1 + (1.0/Q) - 2*sqrt(1/Q));
RMTminIndex = find(eigvals < RMTminEig);
if isempty(RMTminIndex)
    RMTminIndex = 1;
else
    RMTminIndex = RMTminIndex(end);
end

avgEigenValue = mean(eigvals(1:RMTmaxIndex));

Dg = zeros(N,N);

Dg(1 : (N+1) : (RMTmaxIndex-1)*(N+1)) = avgEigenValue;

Dg(1+(N+1)*(RMTmaxIndex-1) : (N+1) : end-(N+1)) = D(1+(N+1)*(RMTmaxIndex-1) : (N+1) : end-(N+1));

M = V * Dg * V.';

M = M - diag(diag(M)) + eye(N);

end
```

Appendix C: PMFG graphs and Newman-Girvin

[1] F. Caravelli et al., our paper.

[2] Girvan M. and Newman M. E. J., Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 7821-7826 (2002)

[3] M. Tumminello, F. Lillo, R.N. Mantegna,Correlation, hierarchies, and networks in financial markets ,J. Econ. Behav. Organ. 75, pp. 40-58 (2010)

[4] S. Gomez, P. Jensen, A. Arenas, Analysis of community structure in networks of correlated data, arXiv:0812.3030

[5] M. MacMahon, D. Garlaschelli, Community Detection for correlation matrices, arXiv:1311.1924

[6] M. Tumminello, T. Aste, T. Di Matteo, R.N. Mantegna, A tool for filtering information in complex systems, Proceedings of the National Academy of Sciences of the United States of America (PNAS) 102 (2005) 10421-10426.

[7] L. Laloux, P. Cizeau, J.-P. Bouchaud, M. Potters, Noise dressing of financial correlation matrices, Phys. Rev. Lett. 83, 1467 (1999)

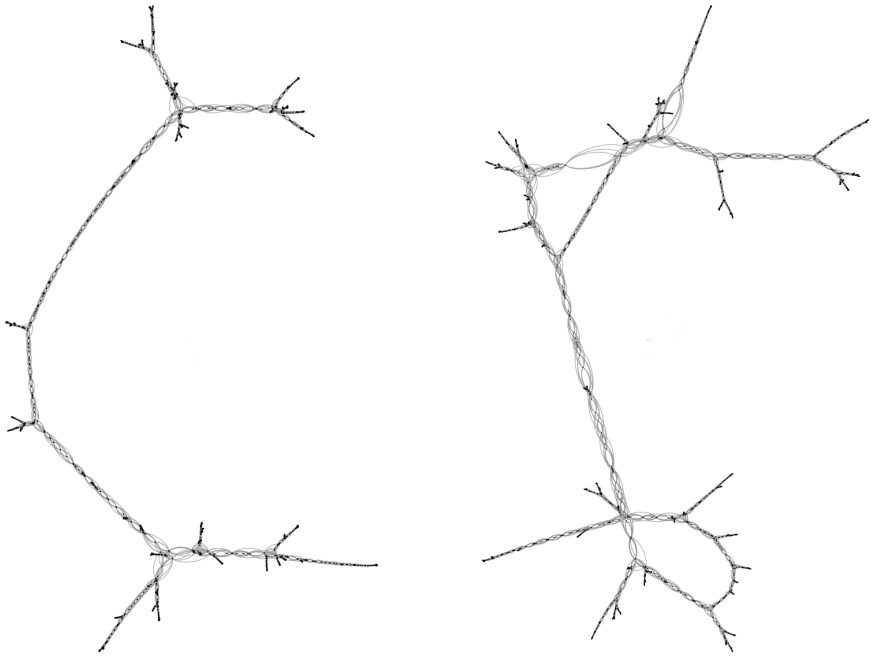

FIG. 9: PMFG based on pearson correlation for PJM (left) and MISO (right).

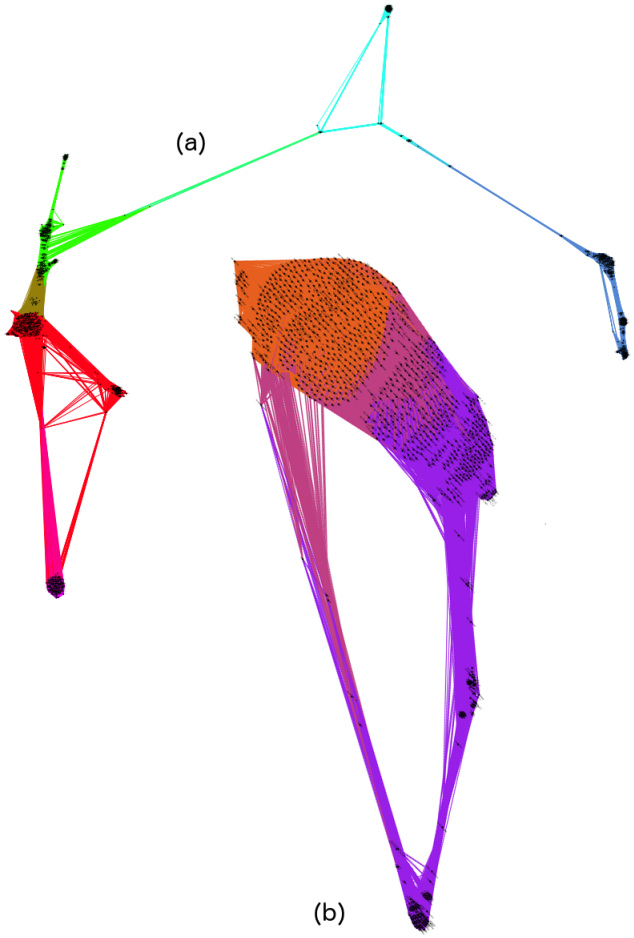

FIG. 10: Newman's clustering algorithm applied to PJM (a) and MISO (b) based on Pearson correlation (where correlations lower than 0.7 have been neglected).