12-30-2008

Topological Structures in the Equities Market Network

Gregory Leibon
Dartmouth College

Scott Pauls
Dartmouth College

Daniel Rockmore
Dartmouth College

Robert Savell
Dartmouth College

Topological structures in the equities market network

Gregory Leibon[a], Scott Pauls[a], Daniel Rockmore[a,b,1], and Robert Savell[c]

Departments of [a]Mathematics and [b]Computer Science and [c]Thayer School of Engineering, Dartmouth College, Hanover, NH 03755

Edited by H. Eugene Stanley, Boston University, Boston, MA, and approved November 5, 2008 (received for review March 20, 2008)

We present a new method for articulating scale-dependent topological descriptions of the network structure inherent in many complex systems. The technique is based on ''partition decoupled null models,'' a new class of null models that incorporate the interaction of clustered partitions into a random model and generalize the Gaussian ensemble. As an application, we analyze a correlation matrix derived from 4 years of close prices of equities in the New York Stock Exchange (NYSE) and National Association of Securities Dealers Automated Quotation (NASDAQ). In this example, we expose (*i*) a natural structure composed of 2 interacting partitions of the market that both agrees with and generalizes standard notions of scale (e.g., sector and industry) and (*ii*) structure in the first partition that is a topological manifestation of a well-known pattern of capital flow called ''sector rotation.'' Our approach gives rise to a natural form of multiresolution analysis of the underlying time series that naturally decomposes the basic data in terms of the effects of the different scales at which it clusters. We support our conclusions and show the robustness of the technique with a successful analysis on a simulated network with an embedded topological structure. The equities market is a prototypical complex system, and we expect that our approach will be of use in understanding a broad class of complex systems in which correlation structures are resident.

cluster analysis | complex systems | multiscale | Delaunay decomposition | dimension reduction

Complex systems often arise as a consequence of multilayered interactions among a large population of diverse agents. For example, neural capabilities arise as a result of the interactions of clusters of neurons of similar function (1). Social networks often function as interacting hierarchies of subnetworks (2, 3), as do link networks for webpages (4). The dynamics of the equities market is driven by interactions among sectors, which are in turn influenced by their component industries and by the strategies of large institutional traders (5). The financial markets are of particular interest for researchers in complex systems, because their intrinsically numerical nature provides a wealth of data for analysis and hypothesis testing. The significant complexity of the web of interdependence in the markets has a natural and informative mathematical formulation in terms of a network encoding the correlation structure of some underlying time series (e.g., price or volume) that measures something of the state of the financial instrument. Indeed, such correlation networks are an important class of networks that fall naturally into the larger class of complex phenomena in which entities in a complex system are related according to some measure of similarity.

In this article, we present a new tool for decomposing these kinds of correlation networks, the ''partition decoupling method.'' It is an iterative method in which spectral considerations (i.e., eigenvalues of a relevant matrix) are used to identify significant clusters via comparison to some relevant random model. The effect of these clusters is removed from the underlying data to reveal a residual layer of interaction ready for another round of structural decomposition. The iterated removal of the ''cluster effect'' is akin (in spirit) to the well-known ''multiresolution analysis'' that accompanies wavelet decompositions in signal and image processing (see e.g. refs. 6 and 7, and ref. 8, which contains a more general view of multiresolution analysis). It is likewise similar to a factor or principal component analysis (9) that creates a succession of approximations to a correlation matrix.

Our approach produces a sequence of partitions of the network, each providing a topological description of an aspect of the network structure. This in turn gives rise to natural hierarchical decompositions of the underlying data stream. The hierarchical structure of the data are also manifested in a multiscale structure in the correlation. The derived partitions suggest a new class of null models introduced herein, the ''partition decoupled null model (PDNM),'' which incorporates the different clusters into a random model. A PDNM is best understood as a generalization of the widely used Gaussian ensemble (GE) null model in which there is a natural incorporation of the structural information associated with the partitions. The PDNM carries with it several interacting partitions, each with its own geometric structure, making it a more textured and potentially more powerful model for comparison. We anticipate that the partition decoupling method (PDM) will be of use in a variety of disciplines in which structure based on similarity measures (e.g., correlation) is expected.

As an example, we give a multipartition analysis of the correlation network of a portion of the equities market. Within each partition, we expose a multiscale network in which nodes at any given scale are aggregations of nodes at a finer scale. The nodes both echo and extend the usual notion of sector in the market. The articulation of topological structure yields our second main result—the unsupervised discovery of nontrivial homology (loops) in the network of clusters, reflecting the well-known phenomenon of capital movement called ''sector rotation.''

Ultimately, we reveal that the equities market may be effectively described as a collection of processes defined on interacting networks—a characterization shared by many diverse complex systems. We demonstrate that by a careful decoupling of network partitions, we may peel apart the layers of network structure to reveal subtle interdependencies among network components and residual network structures hitherto masked by more dominant network processes.

Background. Our approach differs in some important ways from previous applications of clustering techniques to achieve hierarchical decompositions of complex systems—and particularly from previous efforts in the articulation of ''market topology'' as manifested in correlation networks derived from equities. The most important difference is that our model is not strictly hierarchical, but instead details the interaction between a number of different partitions of the network. Our method places no constraint on connectivity of the nodes, whereas purely hierar-

Author contributions: G.L., S.P., D.R., and R.S. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: rockmore@math.dartmouth.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0802806106/DCSupplemental.

chical approaches constrain the complexity (in terms of connectivity) of the defined nodes in some manner [e.g., as a tree (10), with some fixed bound on topological type (11)]. Although recent work (12, 13) emphasizes the need in complex systems analysis for identification of graph partitions with strong clustering properties at multiple scales of interest, the independence assumptions implicit in the single partition clustering solutions tend to obscure subtle intercluster effects revealed by our methodology.

Although our use of the GE null model (see the *Methodology* section for a description of the use of this model and the SI for a brief description of the model itself) as a means of identifying relevant information in our clustering step is in the spirit of random matrix null models (14–17), our method provides a more detailed description of a network by identifying relevant clusters across multiple interacting partitions. We also note that, in contrast to our clustering method, cluster identification using localization of eigenvectors (e.g. ref. 14) generally produces clusters that do not necessarily partition the entire set of equities (however, see ref. 18 for a single partition result).

Methodology

Our methodology is designed to preserve important aspects of system complexity typically lost in the application of dimension reduction techniques. The partition decoupling method is a principled method for generating multipartition descriptions of the system that effectively capture both the dominant structures defining the system and lower order structures that are often obscured by the actions of the dominant processes. It involves combining 2 algorithms, the partition scrubbing method and the hierarchical spectral clustering method.

Partition Scrubbing Method. Beginning with a discrete sample space of nodes or entities $I = \{1, \ldots, N\}$ with associated time series $D = \{D(1), \ldots, D(N)\}$ each of length T, we identify a collection of "characteristic time series V," which capture some aspect of the structure of the D series. Note that these need not be (and rarely are) independent. The idea is that each member of V summarizes some property of the time series in D and projection of D onto the subspace spanned by V yields a dimension reduced representation of D. From this, we then derive a decomposition of D into 2 orthogonal components—the projection of D onto V and a residual component R. The process may then be repeated on R and iteration may be continued until "failure," of which there are 2 types: (*i*) "partitioning failure" occurs when the correlation structure of the residual time series is indistinguishable from the Gaussian ensemble (note that depending on context, this could be replaced by other null models) and we cannot reliably find characteristic times series; and (*ii*) "projection failure" occurs when the characteristic time series are numerically linearly dependent. In the case of projection failure, the projection on V does not have a unique representation in terms of the characteristic time series. Our view is that in each iteration, the removal of the effect of the characteristic time series reveals residual structure that may have been masked by a dominant behavior.

To apply the partition scrubbing method, we derive from D a collection of normalized sequences $D^0(i)$ and given a choice of clustering methods for each $0 \leq \alpha \leq m$ and a collection $D^\alpha(i)$, we produce a mapping $C^\alpha: I \rightarrow \{1, \ldots, |C^\alpha|\}$ where $|C^\alpha|$ denotes the number of clusters generated by the method. We calculate the set of characteristic time series associated with the partition:

$$[V^\alpha_k = \text{mean}\{D^\alpha(i) | C^\alpha(i) = k\}]$$

for $1 \leq k \leq C^\alpha$. Then, $V^\alpha = \{V^\alpha{}_1, \ldots, V^\alpha{}_{|C^\alpha|}\}$. (Note that this method can be generalized to any method of constructing the characteristic time series V.)

Next, we "scrub" the partition to produce $D^{\alpha+1}(i)$ from $D^\alpha(i)$. That is, we decompose $D^\alpha(i)$ into the sum of 2 components: the projection $\mathcal{F}^\alpha(i)$ associated with the clustering C^α and a residual component $\mathcal{R}^\alpha(i)$, so that:

$$D^\alpha = \mathcal{F}^\alpha + \mathcal{R}^\alpha \quad [1]$$

where

$$\mathcal{F}^\alpha(i) = \Pi_{V^\alpha}(\mathcal{D}^\alpha(i)) = \sum_{k=1}^{|C^\alpha|} \tau^\alpha_k(i) V^\alpha_k \quad [2]$$

where Π_{V^α} is the projection onto V^α.

We assume that $\mathcal{R}^\alpha$ is independent of V^α. (Here, "independence" is meant in the statistical sense, namely, that they are not correlated.) Under these assumptions, we can solve for the $\tau^\alpha_k(i)$ via some simple linear algebra.* We call $\tau^\alpha_k(i)$ the "cluster pressure on node i" (at iteration α).

From τ, we create a new collection of "cleaned" time series:

$$[D^{\alpha+1} = \text{norm}(\mathcal{R}^\alpha) = \text{norm}(D^\alpha - \mathcal{F}^\alpha)]$$

with $\text{norm}(\mathcal{R}^\alpha) = (\mathcal{R}^\alpha - \mu^\alpha)/\sigma^\alpha$, where μ^α and σ^α denote the mean and standard deviation of $\mathcal{R}^\alpha$ respectively.

Using this algorithm, each series $D^0(i)$ can be reconstructed from the $D^{m+1}(i)$ from the $\Sigma^m_{\alpha=0}|C^\alpha|$ characteristic time series in $\{V^\alpha\}^m_{\alpha=0}$ and the $\Sigma^m_{\alpha=0}|C^\alpha| + 2(m+1)$ parameters $\{\mu^\alpha(i), \sigma^\alpha(i), \{\tau^\alpha_k(i)\}^{C\alpha}_{k=1}\}^m_{\alpha=0}$ corresponding to the entity. This is our "multiresolution" representation of the original time series data.†

Hierarchical Spectral Clustering Method. To find the partitions needed in the partition scrubbing method, we use an innovative hybrid technique, the hierarchical spectral clustering method (HSCM). This is a principled hierarchical clustering of the correlation network, which proceeds by comparing the eigenvalues of the Laplacian of the correlation network to the eigenvalues of a GE null model associated with the network nodes. The method is suitable for networks in which effects of interest tend to result in stratification of network correlation strengths at particular scales. Given a collection of time series indexed by I, the output of this method is n levels of clusters of the nodes, each of which provides a partition of I.

At the core of the method is the dimension reduction via spectral clustering of a graph Laplacian (19) associated with the correlation matrix. (See the SI for an overview of the method). When presented with a correlation matrix for a sequence of time series, we identify the number of significant clusters and perform spectral clustering. To pick the number of significant clusters, we are guided by the use of the GE null model as a means of determining at what point in the spectrum of the Laplacian we are witnessing a manifestation of random effects. The GE null model, GE(n,m), models n nodes with time series of length m, the entries of which are drawn from independent and identically distributed (i.i.d.) Gaussian random variables. The choice of Gaussian random variables (as opposed to a different distribution) is motivated by our choice of application: The total distribution from the observed data for the equities network is

*We take the inner product of both sides of Eq. **2** with V^α_j for all values of j and solve for $\tau^\alpha_k(i)$. Equivalently: Let $A^\alpha_{k,j} = \text{corr}(V^\alpha_k, V^\alpha_j) \times \text{sd}(V^\alpha_j)$. Let $b(j) = \text{corr}(V^\alpha_j, D^\alpha) \times \text{sd}(D^\alpha(i))$. Solve for $T = (A^\alpha)^{-1} \times b$. Then $\tau^\alpha_k(i) = T(k)$.

†Note that projection failure occurs when the τ^α_k values are not uniquely determined (i.e., the matrix A indicated in footnote * is not invertible). We interpret this as a loss of resolution in the data and/or a build-up of numerical error (and stop iterating if such a failure occurs).

close to Gaussian, with the obligatory fat tails.[‡] We set the number of significant clusters equal to the number of nonzero eigenvalues of our correlation matrix that fall below the minimum of the nonzero eigenvalues of the Laplacian of the correlation matrix associated with each of 100 instances of GE(n,m).

We call this first set of clusters the first "level." To form the remaining levels, we repeat the following 2 steps until we reach a level with < 2 clusters. Given a level j:

1. Form a new correlation matrix Corr(j) by computing the correlations between the mean time series of the clusters of level j.
2. Repeat the comparison to the GE null model and spectral clustering described above to find the $(j + 1)$st level of clusters (i.e., these are clusters of clusters).

The above two steps fail if the comparison to the GE null model yields < 2 significant eigenvalues. This is the partitioning failure we defined above. We call a level "nontrivial" if there is >1 significant eigenvalue.

Partition Decoupling Method (PDM). The PDM consists of the iterative application of the partition scrubbing method using the partitions produced by the HSCM. As a first step, we normalize the series and we set $C^0 \equiv 1$. This is akin to defining a partition with a single characteristic time series V^0 incorporating all nodes. (In our equities example, this corresponds to removing the global market effect by removing the overall daily mean, and is similar to the normalization used in ref. 14.) Then we proceed by using the HSCM to form the partitions needed by the partition scrubbing method. Note that running the HSCM requires choosing a level at each iteration. We express these choices with the partition vector $\langle \mathfrak{l}_1, \ldots, \mathfrak{l}_m \rangle$. A partition vector uniquely determines the PDM's output: the characteristic time series $\{V^{\mathrm{m}}_{\langle \mathfrak{l}_1, \ldots, \mathfrak{l}_m \rangle}\}_{\alpha=0}$ and the constants $\{\mu^{\alpha}_{\langle \mathfrak{l}_1, \ldots, \mathfrak{l}_m \rangle}(i), \ \sigma^{\alpha}_{\langle \mathfrak{l}_1, \ldots, \mathfrak{l}_m \rangle}(i), \{\tau^{\alpha}_k(i)\}^{|C\alpha|}_{k=1}\}^{\mathrm{m}}_{\alpha=0}$ for each entity. Here, $C^{\alpha}_{\langle \mathfrak{l}_1, \ldots, \mathfrak{l}_m \rangle}$ denotes the partition formed during the α iteration of the PDM.

Notice the PDM implicitly defines a restricted class of models via constraints on the covariance structures associated with the traditional GE null model. We refer to such an associated null models as a partition decoupled null model (PDNM). Given a partition vector $\langle \mathfrak{l}_1, \ldots, \mathfrak{l}_m \rangle$, we may construct an associated PDNM by replacing the final D^{m+1} with independent Gaussian random variables and inverting the partition scrubbing method. Notice, if the decomposition terminates with a partitioning failure at the α iteration, then the $D^{\alpha+1}$ time series have a correlation structure that is indistinguishable (in the above spectral sense) from the Gaussian ensemble, and this model duplicates the correlation network structure up to random effects. Decompositions that halt due to a projection failure may still have a residual that has significant structure when compared with a GE null model, but we cannot reliably compute the the $\tau^{\alpha}_k(i)$ that represent the contributions of the clusters.

Decomposition of Equities Networks. For our specific application to the equities market network, we begin with N time series of daily close prices and create an initial collection of series D^0, which corresponds to the so-called "logarithmic return" (approximate logarithmic derivative or fractional change) of the closing price series for each equity,

[‡]As a check, we performed our entire analysis with a bootstrap null model based on the observed data distribution but found no difference (compared with the use of a GE) in the results. Thus, for ease of exposition and replication, we use the Gaussian distribution as our base distribution. For other applications, a different choice of distribution may be appropriate.

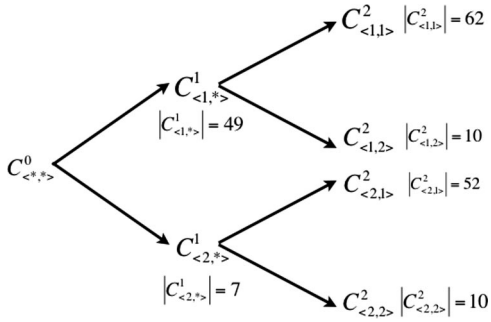

Fig. 1. A tree diagram showing the partitions involved when exploring two iterations of the partition decoupling method with respect the equity market example.

$$D^0_{\mathrm{t}}(0) = \frac{P_{\mathrm{t}}(i) - P_{\mathrm{t}-1}(i)}{P_{\mathrm{t}-1}(i)}$$

where $P_t(i)$ denotes the closing price on day t of equity i.

In this section, we describe the results of applying the PDM to the equities network determined by these series. We demonstrate the ability of the PDM to expose network structures that elude typical clustering methodologies. In doing so, our results delineate a more general notion of market sectors than those typically acknowledged by the industry, in that we expose both recognizable "classical sectors" and new natural hybrids. Additionally, the coarse scale analysis successfully exposes a nontrivial homological entity (a topological cycle) corresponding to the known phenomenology of capital flow referred to as sector rotation.

For this application, we obtained from the *Yahoo!* Finance historical stock data server daily close prices for stocks listed on the New York Stock Exchange (NYSE) and National Association of Securities Dealers Automated Quotation (NASDAQ) during a period spanning 1,251 days of trading between March 15, 2002 and December 29, 2006. We began by removing any equity with $>30\%$ missing data in that window, after which we were left with 2,547 equities. In addition, we remove all extreme events from the time series (20% or larger single-day moves). This cleaning was performed in part to avoid having to carefully compensate for the stock splits and reverse stock splits in our data. However, we feel that this cleaning would be appropriate even if we had cleaned out the splits via other methods. This is because the structure that underlies the market exists in (at least) 2 regimes—extreme events and "normal" events—articulated as 2 different network structures (20). Because the extreme events were very sparse, the time series correlations we used to explore the equities market are by their nature only capable of illuminating the normal network.

PDM Applied to the Equities Market. To demonstrate the method's superiority in exposing latent structure in the network, we look at the results of the PDM for 2 iterations. This resulted in 4 possible partition vectors with nontrivial levels, as schematically described in Fig. 1. We found partitions of the following sizes: $|C^1_{\langle 1,*\rangle}| = 49$, $|C^1_{\langle 2,*\rangle}| = 7$, $|C^2_{\langle 1,1\rangle}| = 62$, $|C^2_{\langle 1,2\rangle}| = 10$, $|C^2_{\langle 2,1\rangle}| = 52$, and $|C^2_{\langle 2,2\rangle}| = 10$. Notice, the partition at the first iteration is independent of future iterations, hence the * can denote any choice. All of these partition vectors provide effective and distinct dimension reductions of the overall complex system.

For each partition, we use the industry sector labelings available from *Yahoo!* Finance and NASDAQ/NYSE membership to examine the composition of clusters as both a validation of our clustering method and a tool that helps show when partitions reveal new information. We find that the majority (35

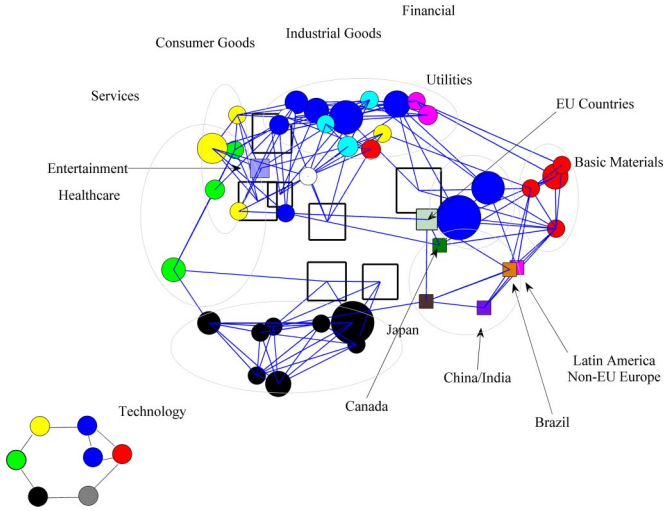

Fig. 2. Network structure after after applying PDM and dimension reduction. The big graph is $C^1_{\langle 1,*\rangle}$ with distance determined by the correlation between the resulting characteristic time series. Solid circular nodes are classified clusters with coloring indicating the dominant sector or classification. Unfilled square nodes are clusters without a dominant classification labeling. Node size in all cases is proportional to cluster size. Connections (blue lines) are added when the Euclidean distance between 2 cluster centroids in the Euclidean embedding is in the bottom 10% of all such distances. The gray ellipses identify clusters of clusters and are (basically) $C^1_{\langle 2,*\rangle}$. A schematic drawing of the resulting network is in the lower left. Nodes are labeled 1–7 counterclockwise beginning with the yellow node.

of 49) of the clusters of $C^1_{\langle 1,*\rangle}$ are predominantly identified by sector (in the sense that the majority of their nodes are from a given sector) and most of the clusters are strongly identified with either the NASDAQ or the NYSE (see SI). Seven of the clusters without dominant sectors have other obvious categorizations (e.g., a regional or business commonality). Clusters of partition $C^1_{\langle 2,*\rangle}$ generally were also classified by sector. Fig. 2 shows a representation of the network resulting from the spectral clustering algorithm applied in our first iteration. For visualization purposes, we have used the centroids of the clusters in $C^1_{\langle 1,*\rangle}$ to represent the entire cluster and have used standard multidimensional scaling (see e.g. ref. 21) to reduce to a lower dimension. The gray regions in Fig. 2 approximately reflect the clusters of $C^1_{\langle 2,*\rangle}$. The graph in Fig. 2 *Inset* shows only the $C^1_{\langle 2,*\rangle}$ clusters and is colored according to dominant sector. Table 1 provides a precise summary of the clustering data and classification. Clusters of $C^2_{\langle 2,*\rangle}$ predominantly admit natural classification (30 of 52 are classified by sector/industry, and 5 more are classified by geography), whereas the opposite is true of clusters of $C^2_{\langle 2,2\rangle}$, where only 3 of 10 admit sector classification (as shown in Fig. 3).

The clusters of $C^2_{\langle 2,2\rangle}$ and $C^2_{\langle 2,1\rangle}$ provide new partitions of the network and reveal new, textured information previously obscured by behavior of the dominant clusters discovered in the first iteration. While clusters of both $C^1_{\langle 1,*\rangle}$ and $C^2_{\langle 2,1\rangle}$ are classified by sector and have significant membership overlap, the network configuration is substantially different from that shown in Fig. 2. This demonstrates that the clusters of $C^2_{\langle 2,1\rangle}$ correspond to a new subsidiary network structure, revealed by exposing new strata of correlation strengths (of lower magnitude) previously masked by the dominant behavior of the clusters in $C^1_{\langle 2,*\rangle}$. For example, although the original cluster of nodes comprising the technology cluster of $C^1_{\langle 2,*\rangle}$ were positively correlated and tightly grouped, the removal of $C^1_{\langle 2,*\rangle}$ via partition decoupling exposes a new configuration for these entities in which there is clustering in similar groupings but with different internal relationships, including negative correlations. It is evident from this analysis that the partition decoupling has removed the major effect of $C^1_{\langle 2,*\rangle}$, revealing lower order effects. We hypothesize that these new partition layers may indicate "second order" trading strategies within these sectors. We note that within the other clusters of $C^1_{\langle 2,*\rangle}$, similar reconfiguration effects are found.

The representation of $C^2_{\langle 2,2\rangle}$ is shown in Fig. 3. The 3 clusters classified by sector reflect reconfigurations of the sectorial

Table 1. Classification of clusters of $C^1_{\langle 1,*\rangle}$ and $C^1_{\langle 2,*\rangle}$

Cluster	Sector	Classification
(1,1)	N	None
(2,6)	N	None
(3,2)	N	None
(4,1)	N	None
(5,1)	N	None
(6,7)	F	Closed end funds
(7,3)	N	None
(8,3)	T	IT Products/services
(9,6)	F	Regional banking, S&Ls
(10,7)	F	Closed-end funds, debt
(11,6)	N	None
(12,1)	S	Strip mall stores
(13,7)	F	REITs
(14,4)	N	EU countries
(15,5)	B	Oil ans Gas
(16,3)	T	Semiconductors, electronics
(17,6)	F	Regional banking, S&Ls
(18,2)	H	Biotechnology
(19,1)	N	Entertainment/leisure
(20,7)	F	Regional banking
(21,3)	T	Software
(22,1)	I	Construction
(23,7)	F	Insurance
(24,2)	H	Drugs/medical supplies
(25,1)	B	Chemicals
(26,7)	U	Electric
(27,5)	B	Industrial metals
(28,3)	T	Scientific/technical Instruments
(29,1)	C	Grocery store items
(30,3)	T	Communication
(31,4)	N	China and India
(32,4)	N	Latin America, non-EU European countries
(33,3)	T	Computer components
(34,1)	S	Media companies
(35,7)	F	Brokerages, asset/credit management
(36,3)	T	None
(37,5)	B	Oil and gas drilling
(38,2)	H	Health care plans
(39,1)	S	Shipping (air and rail)
(40,7)	U	Gas
(41,1)	S	Restaurants
(42,3)	T	Internet services
(43,1)	I	Aerospace Products/services
(44,4)	N	Brazil
(45,1)	C	Auto parts/manufacture
(46,7)	N	Canada
(47,4)	N	Japan
(48,1)	I	Residential construction
(49,5)	B	Gold industries

Clusters are recorded as (a,b) where a is the $C^1_{\langle 1,*\rangle}$ label and b is the $C^1_{\langle 2,*\rangle}$ label. Sectors are identified via *Yahoo! Finance* labels as B (basic materials), C (consumer goods), F (financial), H (healthcare), I (industrial goods), N (none), S (services), T (technology), and U (utilities).

divisions given by $C^1_{\langle 2,2\rangle}$. More interesting are the unclassified clusters that reveal new cross-sector interactions. For example, the diamond shaped clusters contain a mixture of multiple sectors. The first is predominantly consumer goods, industrial goods and services, and the second is predominantly financial, healthcare, services, and technology. However, both clusters contain significant commonalities. In the first, the equities in the service sector are almost all related to the shipping industry, which obviously serves to distribute consumer and industrial goods. In the second, the equities in the financial, services, and technology sectors are related to companies that either provide services or do business with healthcare companies (e.g., health insurance companies, drug companies, management services, healthcare based REITs, etc.). Equities in both of these clusters are drawn from a range of different clusters in $C^1_{\langle 2,*\rangle}$, showing that these two overlapping partitions are truly distinct, and once again demonstrating PDM's ability to remove higher order effects and reveal new structure.

Nontrivial Homology—Sector Rotation. The most significant geometric property of the hierarchical network exposed in the first iteration (as shown in Fig. 2) is the existence of a topological cycle (i.e., an example of nontrivial homology) reflective of the well-known phenomenon of sector rotation—which forms the basis for predictive techniques in intermarket analysis (22). Sector rotation refers to the typical pattern of capital flow from sector to sector over the course of a business cycle. Capital flow is echoed in our network structure via enhanced correlations among related equities, and the topological cycle corresponding to sector rotation manifests itself as an emergent structure in the dense network of near neighbor links. To support the hypothesis that we are exposing sector rotation, we compute the effect of the overall market pressure, τ, for each equity in a moving 1-year window over 10 years of data. Because most of our clusters are sector dominated, we compute, as a proxy for the aggregate pressure on the clusters, the mean τ for each sector.[§] In Fig. 4, we plot the results over time after applying standard normalization. Both the periodicity of the sector effects and the relative phases of the sector waveforms strongly support the sector rotation interpretation.

Visually, the cycle is evident. We support this intuitive conclusion with a Delaunay-type geometric argument in the SI. The method explained therein (see e.g. ref. 23) and our detection of a well-articulated topology in this network is in a similar spirit to the ideas presented in refs. 24 and 25, where the computation of homology of a geometric object based on large datasets is used a means of articulating a topological signature for the dataset. In our case, the homology has a natural interpretation in terms of observed market behavior.

Performance of the PDM on a Simulated Complex System. To demonstrate the effectiveness of the PDM in defining useful dimension reductions of complex system behaviors, we produce a sequence of simulated datasets that manifest a multiscale multipartition correlation structure similar to that found in the equities market. By varying the relative scale of quantized correlation effects producing the simulated time series, we may systematically examine the effectiveness of the PDM relative to a common single partition hierarchical clustering decomposition. As described in detail in the SI, we define a simulated PDNM as a mixture of i.i.d. Gaussian characteristic time series associated to 2 partitions with 7 clusters each. In our experiment, we restrict the degrees of freedom of the simulation and systematically vary the scale (as defined by the τ values associated with each effect) of secondary and tertiary effects relative to the primary effect. We compare the quality of the solution of the PDM on the simulated data with that of a successful spectral clustering method (19) applied to a single partition, by comparing the L^2 deviance of the eigenspectrum of the Laplacian of the GE null model to that of the residual time series resulting from (*i*) a 2-layer application of PDM and (*ii*) an application of the partition scrubbing method using the time series associated with the single partition spectral clustering (with the number of clusters equivalent to the total number comprising the model produced by the PDM).

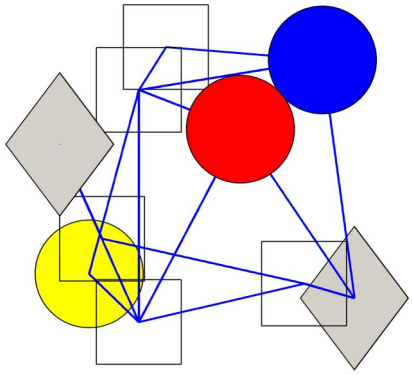

Fig. 3. The network with nodes determined by $C^2_{\langle 2,2\rangle}$ and with distance determined by the correlation between the resulting characteristic time series. Three are identified by sector (red, basic materials; yellow, services; blue, financial). The 2 diamond shaped clusters are classified by intersector commonalities as described in *Methodology*.

By varying the relative proportion of the lower scale effects, we find that when the correlation scales of the two partitions are sufficiently distinct (i.e., the weights of the two sets of characteristic time series are sufficiently disparate), the PDM readily demonstrates its superiority over the single partition spectral clustering in identifying the multipartition structure. Further, in this case the PDM captures more information than the clustering method alone, in the sense that the deviance resulting from the PDM is several orders of magnitude smaller than that of the single-partition model (see SI for details). The PDM maintains this advantage over a broad range of the scaling parameter. Only in cases in which the secondary and tertiary effects are relatively small does the PDM fail to outperform the single partition clustering method, because the second partition goes undetected by the PDM (the HSCM finds no significant eigenvalues as all eigenvalues are above the threshold value). In this case, the results are equivalent to that of the single partition clustering. Hence, as expected, the PDM performs at least and a single partition clustering method across the full range of scale settings and provides superior results when significant scalar quantization effects are present in the correlation structure.

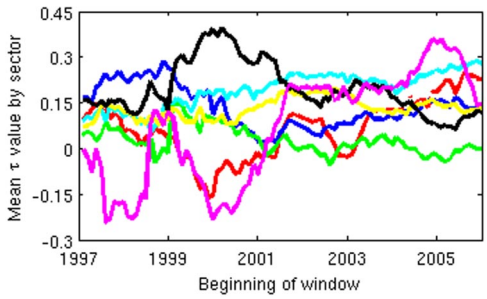

Fig. 4. Average τ by sector of time: 1-year windows over 10 years.

[§]Recall that τ with respect to any subset (including the entire market) is the time series given by the average fractional change over the entire subset on each day.

Having established the ability of the PDM to decouple partitions, we investigated its ability to extract topological structure in the resulting partitions. To do so, we simulated a system with 2 partitions where the clusters in each were designed to produce a topological circle (see SI for details). In these simulations the PDM accurately separated the partitions and preserved the encoded topological structures, as measured by our own implementation (based on ref. 23) of a version of persistent homology (24, 25).

Conclusion

We present a new technique, the partition decoupling method (PDM) for the decomposition of complex systems given a correlation network structure that yields scale-dependent geometric information—which in turn provides a multiscale decomposition of the underlying data elements. The PDM generalizes traditional multiscale clustering methods by exposing multiple partitions of clustered entities.

Our multipartition decomposition allows us to create a new class of null models with which to study such systems, the partition decoupled null model. These null models mimic the observable clustering of the network and thus provide a better platform than the random matrix theory models from which to study the behavior of the network.

As an example and application, we analyze a substantial portion of the U.S. equities market, revealing several partitions that expose 6 different dimension reductions of the market network. Labeling by traditional sector and industry data validate one aspect of the partitioning, as the finest partitions break down both by traditional sector and other commonalities. In addition to validation of the technique (by recovering "official" classifications), the labeling provides evidence for our technique's ability to extend traditional notions of a priori clusters in the data. The partition decoupling reveals several instances of cross-sectorial components (with verifiable mixture classifications), which tend to be obscured by the typical sectorial analysis.

Our decomposition also reveals an instance of nontrivial homology, a cycle that corresponds to the well known phenomena of sector rotation, which reflects the cyclical flow of capital through the the various sectors of the equities market as the economy moves through the stages of expansion and contraction. The visual evidence of the cycle is validated via techniques from computational geometry.

In conclusion, the PDM applied to the correlation network of the equities market reveals both interesting known structure and new structure that is typically lost in common sectorial market decompositions. This principled decomposition of the time series according to the structure of the equities correlation network should prove useful for various forms of risk management including portfolio construction. More generally, we anticipate that correlation networks produced by the actions of other diverse complex systems and other kinds of similarity networks will also prove amenable to this approach.

1. Kandel E, Schwartz J, Jessell T (2000) *Principles of Neural Science* (McGraw–Hill, New York).
2. Watts JD, Muhamad R, Medina DC, Dodds PS (2005) Multiscale, resurgent epidemics in a hierarchical metapopulation model. *Proc Natl Acad Sci USA* 102:11157–11162.
3. Watts DJ, Dodds PS, Newman MEJ (2002) Identity and search in social networks. *Science* 296:1302–1305.
4. Broder A, *et al.* (2000) Graph structure in the Web. *Computer Networks* 33:309–320.
5. Gompers PA, Metrick A (2001) Institutional investors and equity prices. *Q J Econ* 116 (1):229–259.
6. Mallat SG (1989) A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans Pattern Anal Machine Intell* 11:674–693.
7. Barbano PE, Spivak M, Feng J, Antoniotti M, Mishra B (2005) A coherent framework for multiresolution analysis of biological networks with ''memory'': Ras pathway, cell cycle, and immune system. *Proc Natl Acad Sci USA* 102:6245–6250.
8. Rahman I, Drori I, Stodden VC, Donoho DL, Schroeder P (2005) Multiscale representations for manifold valued data. *SIAM J Multiscale Model Sim* 4:1201–1232.
9. Hair J, Black B, Babin B, Anderson RE, Tatham RL (2004) *Multivariate Data Analysis* (Prentice Hall, Englewood Cliffs, NJ).
10. Mantegna NR (1999) Hierarchical structure in financial markets. *Eur Phys J Ser B* 11:193–197.
11. Tumminello M, Aste T, Di Matteo T, RN Mantegna (2005) A tool for filtering information in complex systems. *Proc Natl Acad Sci USA* 102:10421–10426.
12. Sales-Pardo M, Guimera R, Moreira A, Amaral L (2007) Extracting the hierarchical organization of complex systems. *Proc Natl Acad Sci USA* 104:15224–15229.
13. Arenas A, Diaz-Guilera A, Perez-Vicente C (2006) Synchronization reveals topological scales in complex networks. *Phys Rev Lett* 96:114102.
14. Plerou V, et al. (2002) A random matrix approach to financial cross-correlations. *Phys Rev E* 66:066126.
15. Plerou V, Gopikrishnam P, Rosenow B, Amaral L, Stanley HE (1999) Universal and non-universal properties of cross-correlations in financial time series. *Phys Rev Lett* 83:1471.
16. Stanley HE, *et al.* (2002) Self-organized complexity in economics and finance. *Proc Natl Acad Sci USA* 99(Suppl):2561–2565.
17. Laloux L, Cizeau P, Bouchaud JP, Potters M (1999) Noise dressing of financial correlation matrices. *Phys Rev Lett* 83:1467.
18. Kim DH, Jeong H (2005) Systematic analysis of group identification in stock markets. *Phys Rev E* 72:046133.
19. Ng A, Jordan M, Weiss Y (2001) *Advances in Neural Information Processing Systems 14*, eds Dietterich T, Becker S, Ghahramani Z, (MIT Press, Cambridge, MA), pp 849–856.
20. Khandani AE, Lo AW (2007) What happened to the quants in August 2007? Available at http://ssrn.com/abstract=1015987. Accessed October, 2007.
21. Duda RO, Hart PE, Stork DG (1996) *Pattern Classification* (Wiley Interscience, New York).
22. Murphy JJ (2004) *Intermarket Analysis: Profiting from Global Market Relationships.* (John Wiley and Sons, New York) pp 200–215.
23. Fortune S (1992) In *Computing in Euclidean Geometry, Lecture Notes Series on Computing, Volume 1*, eds Du D-Z, Hwang F (World Scientific, Singapore) pp 193–233.
24. Zomorodian A, Carlsson G (2004) In *Proceedings of the Twentieth Annual Symposium on Computational Geometry* (ACM, New York) pp 347–356.
25. Ghrist R (2008) Barcodes: The persistent topology of data. *Bull Am Math Soc* 45:61–75.