A tool for filtering information in complex systems

M. Tumminello*, T. Aste†, T. Di Matteo†, and R. N. Mantegna*‡§

*Istituto Nazionale di Fisica della Materia Unità di Palermo and Dipartimento di Fisica e Tecnologie Relative, Università di Palermo, Viale delle Scienze, I-90128 Palermo, Italy; †Department of Applied Mathematics, Australian National University, Canberra ACT 0200, Australia; and ‡Istituto Nazionale di Fisica Nucleare, Sezione di Catania, I-95125 Catania, Italy

Edited by H. Eugene Stanley, Boston University, Boston, MA, and approved June 10, 2005 (received for review January 13, 2005)

We introduce a technique to filter out complex data sets by extracting a subgraph of representative links. Such a filtering can be tuned up to any desired level by controlling the genus of the resulting graph. We show that this technique is especially suitable for correlation-based graphs, giving filtered graphs that preserve the hierarchical organization of the minimum spanning tree but containing a larger amount of information in their internal structure. In particular in the case of planar filtered graphs (genus equal to 0), triangular loops and four-element cliques are formed. The application of this filtering procedure to 100 stocks in the U.S. equity markets shows that such loops and cliques have important and significant relationships with the market structure and properties.

cluster analysis | complex networks | correlation analysis

Several complex systems have been investigated recently from the perspective of the (weighted) networks that are linking the different elements comprising them (1–4). Indeed, complex systems are in general made of several interacting elements, and it is rather natural to associate to each element a node and to each interaction a link yielding to a graph. Examples include food webs (5), scientific citations (6), social networks (7, 8), communication networks (9), sexual contacts among individuals (10), company links in a stock portfolio (11), the Internet (12), and the World Wide Web (13). The properties of such graphs have been studied with the aim of catching basic features of the investigated systems (14–16). However, the complexity of the system is generally reflected in the associated graph, which results in an intricate interweaved and densely connected structure. There is therefore a general need to find methods that are able to single out the key information by filtering such a complex graph into a simpler relevant subgraph. Such a filtering is especially essential for correlation-based graphs where, in the absence of any filtering procedure, all links among elements are present.

In this work, we introduce a filtering procedure that extracts a representative subgraph with a controlled complexity and maximal information content out of the graph describing the system. To illustrate the method, we present a concrete example dealing with 100 stocks belonging to a U.S. equity portfolio. In the modeling of equity portfolios, a natural starting point is the investigation of cross-correlation among time series of returns of stock pairs. The correlation provides a similarity measure among the behavior of different elements in the system. It was shown by one of us that a powerful method to investigate financial systems consists in the extraction of a minimal set of relevant interactions associated with the strongest correlations belonging to the minimum spanning tree (MST) (11). However, the reduction to a minimal skeleton of links is necessarily very drastic in filtering correlation-based networks, losing therefore valuable information. The necessity of a less drastic filtering procedure already has been raised in the literature. For example, an extension from trees to more general graphs generated by selecting the most correlated links has been proposed in refs. 17–19. However, with the method discussed in refs. 17–19, it is highly improbable to obtain a filtered network connecting all elements via some path by retaining a number of links of the same order as the number of elements.

The method that we present here is based on the key idea that graphs with different degrees of complexity can be constructed by iteratively linking the most strongly connected nodes under the constraint of generating graphs that can be embedded on a surface of a given genus $g = k$ (20). The genus is a topologically invariant property of a surface defined as the largest number of nonisotopic simple closed curves that can be drawn on the surface without separating it, i.e., the number of handles in the surface. We prove that such graphs have the same hierarchical tree associated to the MST (21, 22) but contain a larger amount of information that increases with the genus. We show that, with respect to the MST, the major relative improvement of the information stored in the graph is realized for the planar case when the genus assumes the value $k = 0$.

Filtering Procedure

Construction Algorithm. Let us first illustrate the method and the associated algorithm to filter significant information out of a given complex system composed by n elements where a similarity measure S between pairs of elements is defined, e.g., the weight of links in the original network or the correlation coefficient matrix of the system. An ordered list S_{ord} of pair of nodes can be constructed by arranging them in descending order according to the value of the similarity s_{ij} between elements i and j. Let us first consider the construction algorithm for the MST, as follows: Following the ordered list S_{ord} starting from the couple of elements with larger similarity, one adds an edge between element i and element j if and only if the graph obtained after the edge insertion is still a forest or it is a tree. A forest is a disconnected graph in which any two elements are connected by at most one path, i.e., a disconnected ensemble of trees. With this procedure, the graph obtained after all links of S_{ord} are considered is the MST. In fact, when the last link is included in the graph, the forest reduces to a tree.

In direct analogy with this construction of the MST, we construct graphs by connecting elements with largest similarity under the topological constraint of fixed genus $g = k$. The construction algorithm for such graphs is as follows: Following the ordered list S_{ord} starting from the couple of elements with larger similarity, one adds an edge between element i and element j if and only if the resulting graph can still be embedded on a surface of genus $g \leq k$ after such edge insertion. This algorithm generates simple, undirected, connected graphs embedded on a surface of genus $g = k$. Below, we demonstrate that these graphs have the same hierarchical tree of the MST and that they possess relevant additional information associated with the structure of loops and cliques, making them natural extensions of the MST. A clique of r elements (r-clique) is a complete subgraph that links all r elements.

A special case is when $g = 0$ and the resulting graph is planar (23), i.e., it can be embedded on the sphere. This graph is the first

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: MST, minimum spanning tree; PMFG, Planar Maximally Filtered Graph; r-clique, clique of r elements.

§To whom correspondence should be sent at the * address. E-mail: mantegna@unipa.it.

extension of the MST, and we name it Planar Maximally Filtered Graph (PMFG). An implementation of the algorithm providing the PMFG written in MATHEMATICA is provided as Computer Programs 1 and 2, which are published as supporting information on the PNAS web site. A basic difference of the PMFG with respect to the MST is the number of links, which is $n - 1$ in the MST and $3(n - 2)$ in the PMFG. Conversely, in general, the number of links in a graph G with a fixed genus $g = k$ is at most $3(n - 2 + 2k)$. Therefore, in most practical cases, when $k \sim O(1)$ and $n \gg 1$, the relative increase in the number of links that might be included in the graph by increasing its genus is very small. It follows that the PMFG assumes a special status among all of the graphs constructed with the introduced algorithm. Indeed, it is the simplest and the one providing the most significant additional information with respect to the MST. For this reason, we will give special attention to it. It is worth noting that the construction algorithm and the topological constraints on the PMFG force each element to participate to at least a clique of three elements. In other words, the PMFG is a topological triangulation of the sphere. Only cliques of three and four elements are permitted in the PMFG. Indeed, Kuratowski's theorem (23) does not allow cliques with a number of elements larger than four in a planar graph. Larger cliques can only be present in graphs with genus $k > 0$. The larger the value of k, the larger the number of elements r of the maximal allowed clique [specifically $r \leq (7 + \sqrt{1 + 48k})/2$; see ref. 24].

Hierarchical Organization. We prove the following statement: At any step of construction of the MST and graph G of genus $g = k$, if two elements are connected via at least one path in one of the considered graphs, then they also are connected in the other one. To this end, we must recall the concept of bridge: a link between two elements is a bridge whenever the elements are disconnected via any path in its absence. It follows from the definition of MST that all links in the MST are bridges. Conversely, for graphs with a fixed genus, we have the following important property: If a bridge is inserted between two previously unconnected regions of a graph G, characterized by the genus $g = k$, then the genus of the graph obtained after the insertion is still k. This property is straightforwardly proved as a corollary of the Miller theorem (25, 26) by noting that the addition of a bridge to a graph leaves unchanged the biconnected components of the graph. The above property implies that if the construction algorithm of G selects a link that is a bridge for the graph at that step of construction, then the link is always added to the graph. We now prove the above statement by induction. In the following, we indicate as MST_m and G_m the graphs constructed by using the similarity measure up to the mth row of S_{ord}. For the first two steps of construction, the statement is true: MST_2 and G_2 graphs are always equal. Now suppose the statement is true at the step m of construction, i.e., for G_{m} and MST_m. For the step $m + 1$, only four cases are possible:

(*i*) The new link, connecting the vertices i and j, is a bridge for the MST_{m+1}. By the definition of bridge, this statement implies that the vertices i and j are not connected via any path in MST_m. Therefore, by inductive hypothesis, the vertices i and j are not connected via any path also in G_m, and then the new link is a bridge for G_{m+1} too. In this case, both graphs will include the considered link, and then the statement is true at the step $m + 1$.

(*ii*) The link is a bridge for G_{m+1}. By using the same reasoning as in (*i*), this statement implies that the same link must also be a bridge for the MST_{m+1} due to the inductive hypothesis, and both graphs will include the considered link, and then the statement is true at the step $m + 1$. In the remaining two cases, we assume the condition that the link between the vertices i and j is not a bridge for both MST_{m+1} and G_{m+1}. This condition can be used without loss of generality because if the link is not a bridge for MST_{m+1} (or G_{m+1}), then one always concludes that the link is also not a bridge for G_{m+1} (or MST_{m+1}) by following the same reasoning of case (*i*).

(*iii*) The link is not a bridge for both MST_{m+1} and G_{m+1}, and the genus condition $g \leq k$ fails. In this case, the link is not included to any of both graphs, and the statement is again true at the step $m + 1$.

(*iv*) The link under investigation is not a bridge for both MST_{m+1} and G_{m+1}, and the genus condition $g \leq k$ is satisfied. In this case, G_{m+1} includes the link and MST_{m+1} does not. However because the link added to G_m is not a bridge, the connectivity between pairs of elements in MST_{m+1} and in G_{m+1} rests unchanged in both MST_m and G_m, and the statement is also verified in this last case.

The statement is therefore true, and it has an important implication: The fact that the MST is formed only by bridges implies that the MST is always contained in any graph G of genus $g = k$ generated with the construction algorithm presented above and, as a specific case, in the PMFG. Moreover, this statement shows an even more important fact: The formation of connected clusters of nodes in G_m coincides with the formation of the same clusters in the MST_m. In other words, the hierarchical tree associated with graph G coincides with that of the MST. It is worth noting that the construction algorithm and the associated network properties also hold true in the more general case of weighted networks and non-fully connected networks. In other words, the algorithm is general, and in the case of a nonconnected graph the filtered graph G of genus $g = k$ also will be a nonconnected graph, whereas the equivalent of MST will not be a tree but a forest.

An Illustrative Example. We present here a simple example showing how the PMFG provides additional information with respect to the one contained in the MST and in the associated hierarchical tree. Let us consider a simple system composed by 10

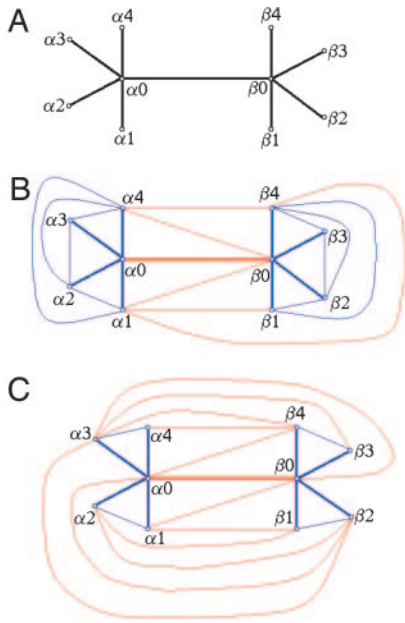

Fig. 1. An illustrative example of two graphs that share the same MST but have distinct PMFGs. (*A*) MST of a simple system of 10 vertices. (*B* and *C*) PMFG of two systems with the same MST (the one drawn in *A*). The thicker lines are identifying links belonging to both the MST and the PMFG, whereas the thinner lines belong to the PMFG only.

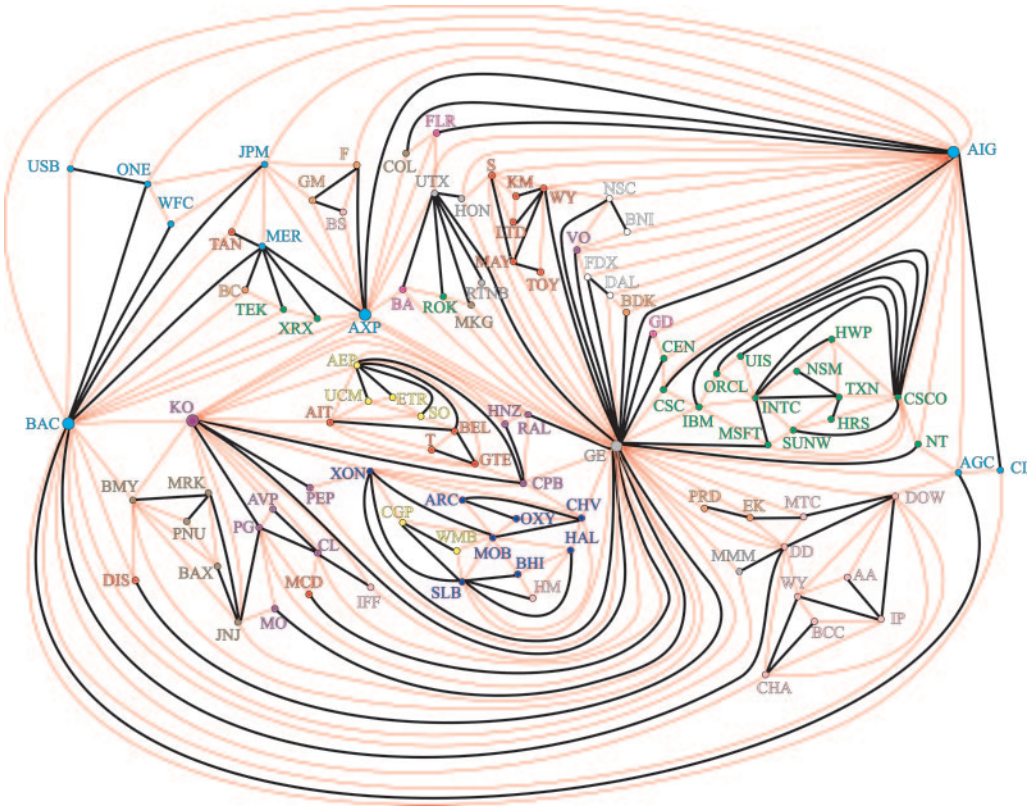

Fig. 2. PMFG obtained from the fully connected graph associated with the correlation coefficient matrix of the 100 most capitalized stocks traded in the U.S. equity markets during the time period from 1995 to 1998. Cross-correlation is computed by using daily returns of stocks. Stocks are indicated with their ticker symbols. For information about a specific ticker symbol, see Table 7, which is published as supporting information on the PNAS web site. The graph is topologically planar: It can be drawn on the plane without edge-crossings. The thicker lines belong to the associated MST. It should be noted that link lengths do not reflect the value of the similarity measure between vertices.

elements being characterized by the MST shown in Fig. 1*A*. Not all of the information about the similarity measure is used to obtain the MST, and different similarity matrices can have the same MST and hierarchical tree but end up into different PMFGs. Two quite distinct possible PMFGs with the same underlying MST are shown in Fig. 1 *B* and *C*. In this figure, the thicker lines are identifying links belonging to both the MST and the PMFG, whereas the thinner lines belong to the PMFG only. By looking just at the MST, we observe in both examples of Fig. 1 *B* and *C* two groups, specifically the one formed by vertices αi ($i = 0, \ldots, 4$), say cluster α, and the other composed by vertices βi ($i = 0, \ldots, 4$), say cluster β. The two clusters are connected by the link between $\alpha 0$ and $\beta 0$. The differences between the two PMFGs can be quantified by a comparison between the link structures of these two planar graphs. In the first planar graph (Fig. 1*B*), there are 18 intracluster links (connecting elements αi to αj and βi to βj) and 6 intercluster links (connecting elements αi to βj). In the second planar graph (Fig. 1*C*), the number of intracluster links is 12, equal to the number of intercluster links. Another way to quantify the differences between the two PMFGs is by analyzing the clique structure of the two planar graphs. In the graph of Fig. 1*B*, a total of seven cliques of four elements is found, with two of them, $\{\alpha 0, \alpha 1, \alpha 2, \alpha 4\}$ and $\{\alpha 0, \alpha 2, \alpha 3, \alpha 4\}$, being characterized by elements all belonging to the cluster α, the other two having elements internal to the cluster β, $\{\beta 0, \beta 1, \beta 2, \beta 4\}$ and $\{\beta 0, \beta 2, \beta 3, \beta 4\}$, and three being composed of intercluster elements, $\{\alpha 1, \alpha 4, \beta 0, \beta 4\}$, $\{\alpha 1, \beta 0, \beta 1, \beta 4\}$, and $(\alpha 0, \alpha 1, \alpha 4, \beta 0\}$. Looking at the cliques of 3 elements, we observe 14 intracluster and 8 intercluster cliques. Quite differently, in the graph of Fig. 1*C* we observe just 4 cliques of 4 elements, no one of them being an intracluster clique, and just 4 of the 20 cliques of 3 elements are intracluster cliques. This analysis shows that the two considered PMFGs refine in a quite different way the information associated with the (same) MST. This result shows that the PMFG provides additional information with respect to the MST, and this information is primarily associated with the nature of the observed cliques.

An Empirical Application

PMFG for 100 U.S. Stocks. Above, we introduced a general method for constructing graphs of defined genus $g = k$ by using a similarity measure. In the following, we constrain ourselves to the case $g = 0$, i.e., to the PMFG, and we present an example concerning the filtering of a graph obtained with a correlation-based procedure. Specifically, we consider pair correlation between daily returns of a set of 100 stocks traded in the New York equity markets in the time period between January 1995 and December 1998 (27). In this case, the measure of similarity S is given by the correlation coefficient ρ_{ij} between stocks. The PMFG for this system is shown in Fig. 2 (see also Fig. 4, which is published as supporting information on the PNAS web site). In Fig. 2, the various elements are connected through links, with thicker lines indicating links belonging to both the MST and the PMFG. By comparing the PMFG and the MST, several new details emerge, with the most striking difference being that the PMFG allows the existence of loops and cliques as we saw in the above example. In the PMFG, we counted $292 = 3n - 8$ cliques of 3 elements out of the possible $\binom{n}{3} = 161{,}700$. The number of cliques of 4 elements is $97 = n - 3$. This number is much smaller than the number of possible cliques of 4 elements present in the

Table 1. Strongest correlated intrasector 4-cliques

Economic sector — Intrasector 4-cliques

Sec.	c_4	Stock 1	Stock 2	Stock 3	Stock 4	$\langle\rho\rangle$	$\langle y\rangle$
E	5	ARC	CHV	MOB	XON	0.628	0.335
B	4	BCC	CHA	IP	WY	0.592	0.334
F	6	AXP	BAC	JPM	MER	0.589	0.334
T	8	CSCO	INTC	MSFT	SUNW	0.537	0.335
H	2	BAX	BMY	JNJ	MRK	0.465	0.339
C	2	AVP	CL	KO	PG	0.462	0.337
S	3	AIT	BEL	GTE	T	0.422	0.354
U	1	AEP	ETR	SO	UCM	0.398	0.343

Table 2. 4-Cliques belonging to the Technology sector

Stock 1	Stock 2	Stock 3	Stock 4	$\langle\rho\rangle$	$\langle y\rangle$
CSCO	INTC	MSFT	SUNW	0.537	0.335
CSCO	IBM	INTC	MSFT	0.534	0.335
CSCO	INTC	SUNW	TXN	0.519	0.335
CSCO	HWP	INTC	TXN	0.503	0.336
CSCO	IBM	INTC	ORCL	0.475	0.336
HWP	INTC	NSM	TXN	0.471	0.339
CSCO	HRS	SUNW	TXN	0.435	0.338
CSCO	INTC	ORCL	UIS	0.380	0.354

fully connected graph, which is $\binom{n}{4} \cong 3.92 \times 10^6$. The complete lists of cliques with three and four elements present in the PMFG are provided in Tables 5 and 6, which are published as supporting information on the PNAS web site. Interestingly, these numbers of three- and four-element cliques coincide with the numbers of such cliques attainable when a graph is made by a set of tetrahedra (4-cliques) packed together by sharing a triangular face. The fact that we observe such numbers of cliques can be qualitatively explained. Consider three elements of a correlation based network, say A, B, and C. If A is strongly correlated to B, and B is strongly correlated to C, then a strong correlation between A and C also should be detected, which makes highly probable the formation of a triangular clique. Now, if one of these three elements is strongly correlated with a fourth one, say D, then also the other two are likely to have a strong correlation with D generating in this way a 4-clique: a tetrahedron. Given the topological constraint of planarity, the next most correlated element can only be connected to maximum three of the four elements of such tetrahedron. The connection of a new element to three elements of the 4-clique generates another 4-clique, which is a new tetrahedron sharing a face with the previous one. By following this reasoning, we expect therefore that the basic structures in the resulting graph are the 4-cliques, which during the formation of the PMFG clusterize together locally at similar correlation values and then connect to each other by following the MST as skeleton structure. Therefore, if such four-element cliques are the "building blocks" of the PMFG, then there must be strong relations between their properties and those of the system of 100 stocks from which they have been generated. These relations are explored below.

Financial Market Properties and 4-Cliques Structure. Let us first classify each stock accordingly with an economic sector following the classification of Forbes Magazine. An analysis on all of the 4-cliques in the PMFG reveals a high degree of homogeneity with respect to the economic sectors. Indeed, we observe that 31 of the 97 cliques are composed by stocks belonging to the same economic sector, 22 are composed by 3 stocks belonging to the same sector, 37 have 2 stocks from the same sector, and only 7 have stocks all from different sectors.

In Table 1, we list the eight cliques with the largest mean correlation $\langle\rho\rangle$ among stocks for each economic sector having at least one clique of four elements. We label the economic sectors as Energy (E), Basic Materials (B), Financial (F), Technology (T), Healthcare (H), Consumer Noncyclical (C), Services (S), and Utilities (U). The total number of intrasector 4-cliques (c_4) for each sector is given in the second column of the table. It should be noticed that $\langle\rho\rangle$ among stocks is different for different sectors. For example the clique with the largest mean correlation is a clique of the Energy sector, which has $\langle\rho\rangle = 0.628$, whereas the clique of the sector Utilities has the smallest mean correlation with $\langle\rho\rangle = 0.398$. To better understand the structure of such cliques, it is interesting to quantify how much the correlation among the stocks is spread within the clique. In analogy to refs. 28 and 29, we compute the quantity $\langle y\rangle$ inside a clique as the mean value of the disparity measure

$$y(i) = \sum_{j\neq i, j\in \text{clique}} \left[\frac{\rho_{ij}}{s_i}\right]^2,$$

over the clique, where i is a generic element of the clique and

$$s_i = \sum_{j\neq i, j\in \text{clique}} \rho_{ij},$$

is the strength of the element i. This definition is meaningful if $\rho_{ij} \geq 0$ as in the case considered. The value of the disparity is expected to be close to 1/3 for 4-cliques characterized by links with comparable values of the similarity measure. An inspection of the last column of Table 1 shows that most of the cliques have a disparity measure very close to 1/3. Exceptions are the cliques of the sectors Services and Utilities that have a slightly smaller homogeneity in the pair correlation between stocks belonging to the cliques. In Table 2, we present all of the eight cliques of four elements observed for stocks belonging to the Technology sector. Note that also inside a single sector, the level of correlation of the selected cliques may significantly vary. In fact, it ranges from $\langle\rho\rangle = 0.380$ to $\langle\rho\rangle = 0.537$, showing that the PMFG is able to select cliques at different levels of correlation. The selection among all of the possible cliques present in the fully connected graph is rather severe; in fact, for the Technology sector we have 17 elements, and therefore the number of cliques of 4 elements all belonging to this sector that are present in the fully connected graph is 2,380. In other words, only 8 of the possible 2,380 cliques of 4 elements of the fully connected graph are selected by the PMFG.

To elucidate the type and amount of information gained by extending the graph from the MST to the PMFG, we focus in more detail on the intrasector and intersector cliques found by the PMFG and, of course, absent in the MST. From Table 3 one notes that no 4-cliques are observed for the sectors Conglom-

Table 3. Intrasector connection strength

Sec	n_s	$c_4/[n_s - 3]$	$c_3/[3\ n_s - 8]$
E	8	5/5 = 1	16/16 = 1
F	10	6/7 ≅ 0.86	20/22 ≅ 0.91
T	17	8/14 ≅ 0.57	26/43 ≅ 0.60
B	11	4/8 = 0.5	14/25 = 0.56
H	7	2/4 = 0.5	7/13 ≅ 0.54
U	6	1/3 ≅ 0.33	4/10 = 0.40
S	13	3/10 = 0.3	12/31 ≅ 0.39
C	10	2/7 ≅ 0.29	8/22 ≅ 0.36
CO	5	0/2 = 0	2/7 ≅ 0.29
CC	6	0/3 = 0	0/10 = 0
TR	4	0/1 = 0	0/4 = 0
CG	3	—	0/1 = 0

Table 4. Intersector 4-cliques connecting 4 different sectors

Stock 1	Stock 2	Stock 3	Stock 4	$\langle\rho\rangle$	$\langle y\rangle$
BAC (F)	BMY (H)	GE (CO)	KO (C)	0.483	0.336
BAC (F)	BMY (H)	GE (CO)	DIS (S)	0.435	0.337
AIG (F)	GE (CO)	NSC (TR)	WMT (S)	0.423	0.339
AIG (F)	GE (CO)	NSC (TR)	VO (C)	0.400	0.340
AIG (F)	GE (CO)	FDX (TR)	VO (C)	0.374	0.345
AIG (F)	BDK (CC)	DAL (TR)	GE (CO)	0.360	0.346
AIG (F)	CEN (T)	GD (CG)	GE (CO)	0.340	0.351

erates (CO) composed by five stocks ($n_s = 5$), Consumer cyclical (CC, $n_s = 6$), and Transportation (TR, $n_s = 4$), even if the number of stocks n_s composing the sectors would potentially allow them. It also should be noted that the sector Capital Goods (CG, $n_s = 3$) does not form a 3-clique. This observation shows that the PMFG conveys information that is not directly present in the market classification of Forbes or in the MST. In fact, this information is associated with the clustering strength of economic sectors that have been detected by the MST.

To quantify the degree of connection strength of elements, we propose to consider the ratio between the number of 4-cliques (c_4) and 3-cliques (c_3) present among n_s stocks belonging to a given set and a normalizing quantity. These normalizing quantities are $n_s - 3$ for 4-cliques and $3n_s - 8$ for 3-cliques. Although we lack a formal proof, our investigations suggest that these numbers are the maximal number of 4-cliques and 3-cliques, respectively, that can be observed in a PMFG of n_s elements.

In Table 3, the connection strength is presented for all of the elements belonging to the economic sectors both for 4-cliques and 3-cliques. Table 3 clearly shows that the connection strength can be quite different across sets of elements. Specifically, elements belonging to some economic sectors are strongly connected within themselves, whereas others are much less connected. Examples of strong connection are the elements of the Energy and Financial sectors, whereas elements belonging to the Conglomerates, Consumer cyclical, Transportation, and Capital Goods sectors are weakly connected.

In Table 4, we list the seven cliques with all of the four components belonging to different economic sectors. These 4-cliques provide bridging regions among areas of the PMFG populated by elements belonging to different economic sectors. This interpretation is supported by the fact that in these cliques, some of the most connected stocks are present. In fact, General Electric (GE) is present in all of the seven cliques, whereas the American International Group (AIG) is present in five of them.

A joint reading of Tables 3 and 4 shows that economic sectors are not equivalent in the detected PMFG. Specifically, the Energy, Technology, and Basic Materials sectors are sectors of elements significantly connected among them but weakly interacting with stocks belonging to different economic sectors. Indeed, no Energy and Basic Materials stocks and only one Technology stock (CEN) appear in Table 4. Quite differently, the Financial sector (F) has still elements strongly connected among them, but it also participates to all of the seven intersector cliques of Table 4. In other words, various sets of elements may or may not be clustered among themselves and may or may not also be connected to elements of other sectors. The PMFG is able to extract this information. Finally, elements of sectors CO and TR are very weakly clustered among them but are often present in cliques of Table 4 meaning that stocks belonging to these sectors behave like bridges between the Financial, Technology, Healthcare, Services, and Consumer noncyclical sectors. In summary, Tables 3 and 4 show how the PMFG is able to quantify the connection strength of elements of the graph through an analysis of the clique structure, information that is only partially present in the MST and the hierarchical tree.

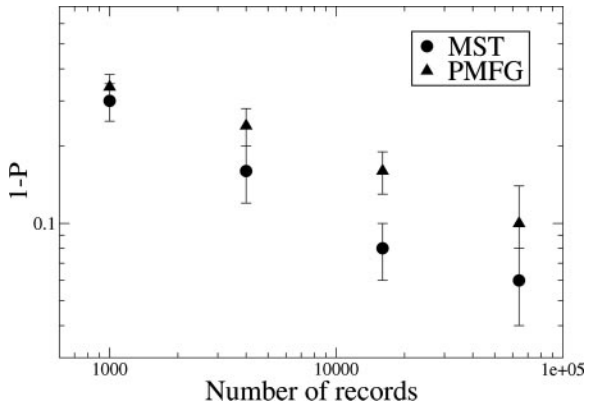

Fig. 3. Analysis of stability of the MST and PMFG with respect to the statistical uncertainty present in the estimation of the correlation matrix as a function of the number of records of the multivariate time series. By assuming as reference matrix the empirical correlation matrix of the system, we perform 4 sets of 20 realizations, each one of surrogated multivariate time series. Each set is characterized by a different number of records set as follows: 1,000, 4,000, 16,000, and 64,000 records. For each of the simulated realizations, both the MST and PMFG have been constructed. The percent ($1 - P$) of the number of links of the simulated graphs nonmatching with the links of the MST and PMFG of real data are shown as a function of the number of records of the surrogated time series in a log–log plot. (The error bar indicates one standard deviation of $1 - P$ computed for each set.)

The analyzed correlation structure has a certain degree of statistical uncertainty because of the finite length of time series. The stability of the filtered graphs with respect to such statistical uncertainty has been analyzed by generating surrogated data series using the discrete Karhunen–Loève expansion (30). The random multivariate Gaussian data series are computed starting from a given correlation matrix. For any simulated realization, the correlation matrix has been calculated. The computed matrices become closer and closer to the reference matrix by increasing the number of records of the simulated time series. We consider the empirical correlation matrix associated to the system as the reference matrix. For fixed values of the number of records of time series, 20 realizations are simulated, and for each of them both the MST and the PMFG have been determined. In Fig. 3, the percent of nonmatching edges in the simulated and the real data graphs is plotted as a function of the number of records of the simulated time series in a log–log scale. Fig. 3 shows that the MST is marginally more stable than the PMFG. Fig. 3 also suggests a power law dependence of the stability of the MST and PMFG with respect to the number of data in the multivariate time series. The significant increase of information gained by the PMFG is therefore fully balancing the marginal decrease of stability for any number of records of the multivariate time series. This finding is another reason suggesting that the PMFG and the similar graphs characterized by a low value of the genus are the best compromise, allowing one to consider a graph richer than the MST but characterized by a similar degree of stability with respect to the statistical uncertainty unavoidably associated with graphs modeling complex systems.

Conclusions

In summary, we have shown that it is possible to determine a family of graphs having the same hierarchical tree associated to the MST but comprising a larger number of links and allowing closed loops. The amount of filtered information with respect to

that present in the MST increases by increasing the genus. A substantial amount of additional filtered information is already obtained in the case for genus $k = 0$, which gives the PMFG. A correlation-based investigation of a financial portfolio shows that the method is pretty efficient in filtering relevant information about the connection structure both of the whole system and within each cluster. The example presented is representative of a large class of correlation-based clustering investigations. For this example, the stability of MST and PMFG with respect to the statistical uncertainty due to the finite length of time series has been investigated, suggesting a power law dependence of the stability with respect to the number of records in time series. The stability of MST turned out to be slightly higher than that of PMFG. The proposed procedure can be applied to a large number of real and correlation-based networks when a filtering of the similarity measure matrix is needed.

We thank S. T. Hyde for fruitful discussions and advice. This work was partly supported by Ministero dell'Istruzione, dell'Università e della Ricerca Research Project 449/97, titled "Dinamica di Altissima Frequenza nei Mercati Finanziari." M.T. and R.N.M. were supported in part by Ministero dell'Istruzione, dell'Università e della Ricerca–Fondo per gli Investimenti della Ricerca di Base Research Project RBNE01CW3M and European Union New and Emerging Science and Technology Project 012911, titled "Human Behavior Through Dynamics of Complex Social Networks: An Interdisciplinary Approach." T.A. and T.D.M. were supported in part by Australian Research Council Discovery Projects DP0344004 (2003) and DP0558183 (2005) and the Australian Partnership for Advanced Computing National Facilities (APAC).

1. Strogatz, S., H. (2001) *Nature* **410,** 268–276.
2. Albert, R. & Barabási, A., L. (2002) *Rev. Mod. Phys*. **74**, 47–97.
3. Dorogovtsev, S. N. & Mendes, J. F. F. (2002) *Adv. Phys.* **51,** 1079–1187.
4. Newman, M., E., J. (2003) *SIAM Rev.* **45,** 167–256.
5. Garlaschelli, D., Caldarelli, G. & Pietronero, L. (2003) *Nature* **423,** 165–168.
6. Redner, S. (1998) *Eur. Phys. J. B* **4,** 131–134.
7. Wasserman, S. & Faust, K. (1994) *Social Network Analysis: Methods and Applications* (Cambridge Univ. Press, Cambridge, U.K.).
8. Newman, M. E. J., Watts, D. J. & Strogatz, S. H. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 2566–2572.
9. Pastor-Satorras, R. & Vespignani, A. (2001) *Phys. Rev. Lett.* **86,** 3200–3203.
10. Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E. & Aberg, Y. (2001) *Nature* **411,** 907–908.
11. Mantegna, R. N. (1999) *Eur. Phys. J. B* **11,** 193–197.
12. Faloutsos, M., Faloutsos, P. & Faloutsos, C. (1999) *Comput. Commun. Rev.* **29,** 251–262.
13. Albert, R., Jeong, H. & Barabási A. L. (1999) *Nature* **401**, 130–131.
14. Watts, D. J. & Strogatz, S. H. (1998) *Nature* **393,** 440–442.
15. Barabási A. L. & Albert, R. (1999) *Science* **286,** 509–512.
16. Amaral, L. A. N., Scala, A., Barthélemy, M. & Stanley H. E. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 11149–11152.
17. Onnela, J.-P., Chakraborti, A., Kaski, K. & Kertész, J. (2002) *Eur. Phys. J. B* **30,** 285–288.
18. Onnela, J.-P., Chakraborti, A., Kaski, K., Kertész, J. & Kanto, A. (2003) *Phys. Rev. E* **68**, 056110.
19. Onnela, J.-P., Chakraborti, A., Kaski, K. & Kertész, J. (2003) *Physica A* **324,** 247–252.
20. Aste, T., Di Matteo, T. & Hyde, S. T. (2005) *Physica A* **346**, 20–26.
21. Gower, J. C. & Ross, G. J. S. (1969) *Appl. Stat.* **18,** 54–64.
22. Mantegna, R. N. & Stanley, H. E. (2000) *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge Univ. Press, Cambridge, U.K.), pp. 120–128.
23. West, D. B. (2001) *An Introduction to Graph Theory* (Prentice–Hall, Englewood Cliffs, NJ), pp. 247–281.
24. Ringel, G. (1974) in *Map Color Theorem* (Springer, Berlin), Chap. 4.
25. Miller, G. L. (1985) *An Additivity Theorem for the Genus of a Graph*, TR-85-342 (MIT, Cambridge, MA).
26. Miller, G. L. (1987) *J. Combin. Theory B* **43,** 25–47.
27. Bonanno, G., Lillo, F. & Mantegna, R. N. (2001) *Quant. Finance* **1**, 96–104.
28. Barthélemy, M., Barrat, A., Pastor-Satorras, R. & Vespignani, A. (2005) *Physica A* **345,** 34–43.
29. Barrat, A., Barthélemy, M., Pastor-Satorras, R. & Vespignani, A. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 3747–3752.
30. Johnson, G., E. (1994) *Proc. IEEE* **82,** 270–285.