Extracting the correlation structure by means of planar embedding

T. Di Matteo and T. Aste

Department of Applied Mathematics, Research School of Physical Sciences, The Australian National University, 0200 Canberra, ACT, Australia.

ABSTRACT

The hierarchical structure of correlation matrices in complex systems is studied by extracting a significant sub-set of correlations resulting in a planar graph. Such a graph has been generated by a method introduced in Aste et al.[1] and it has the same hierarchical structure of the Minimum Spanning Tree but it contains a larger amount of links, loops and cliques. In Tumminello et al.,[2] we have shown that this method, applied to a financial portfolio of 100 stocks in the USA equity markets, is pretty efficient in filtering relevant information about the system clustering revaling the hierarchical organization in the whole system and within each cluster. Here we discuss this filtering correlation procedure and its application to different financial data sets.

Keywords: Econophysics, Complex systems, Networks

1. INTRODUCTION

The correlation matrix is a powerful instrument to describe the collective behavior of systems comprised of many elements. However in many practical cases it turns out that the information in the whole correlation matrix is too detailed, making difficult to extract the essential features associated with the collective properties of the system. Indeed, correlation coefficients are associated to all the pairs of elements in the system. The challenge is to extract the smallest sub-set of such correlations which is able to describe accurately the collective behavior of the whole system. The correlation matrix can be represented -topologically- by the complete graph (each node connected with all the other nodes) with edges weighted by the value of the correlation coefficient. The problem of extracting a system of meaningful interactions can be therefore translated into the problem of reducing the complete graph into a sub-graph which keeps only the edges which best describe the system of interactions. Such a sub-graph must contain the maximum amount of information about the system's collective behavior while keeping the simplest possible structure. It was shown in Ref.[3] that a method to investigate correlations in financial systems consists in extracting a minimal set of relevant interactions associated with the strongest correlations belonging to the Minimum Spanning Tree (MST). The MST maintains only the minimum number of links necessary to connect the network ($n - 1$ links for a network with n nodes) and, by construction, it keeps the links associated with the strongest correlations. This method turns out to be very effective in revealing the hierarchical structure contained in the correlation matrix. However, the reduction to a minimal skeleton of links is necessarily very drastic in filtering correlation based networks loosing therefore valuable information.

Since the MST method has been proved to be very effective in extracting hierarchies from the correlation matrix, its extension to a richer graph must be constructed in a way to maintain such a hierarchical skeleton while including a larger amount of links. A solution to this problem is a new technique which enables to produce networks with tunable information content.[1,2,4] This method is based on the idea of connecting iteratively the most correlated nodes while constraining the resulting network to be embedded on a given surface.

We here show some of the results of the application of this filtering procedure to the correlation matrix of *16 Eurodollar Interest Rates*,[5] 34 *different kinds of interest rates in money and capital markets*[4,6] and 100 US stocks.[2]

The paper is organized as following: in Section 2, we present the general idea regarding the embedding of the complete graph on hyperbolic surfaces and we discuss the case of planar embeddings by considering the Planar

---
Corresponding author: Tiziana Di Matteo, e-mail: tiziana.dimatteo@anu.edu.au; phone 61 2 6125 0166; fax 61 2 6125 0732

Maximally Filtered Graph (PMFG) introduced in Ref..[1,2] Section 3 shows results for different sets of financial data and it analyzes and discusses the emergent clique structure.

2. NETWORKS RESULTING FROM THE CORRELATION MATRIX

Consider a system of n interacting elements which are collectively fluctuating in a stochastic way. In first place, these mutually-related stochastic variables can be considered all connected to each other in an n-th order complete graph (K_n). A *weight* can be associated to each link and a natural choice for the weight of a link between node i and node j is the correlation coefficient $c_{i,j}$. The challenge is to locally simplify the network by keeping only the most significant interactions (largest correlations) and simultaneously extracting global information about the hierarchical organization of the whole system. Hereafter we show that a very powerful way to proceed is to map the complete graph into a significant sub network on a 2-dimensional (2D) hyperbolic surface. This approach has several attractive features: 1) it provides new measures to characterize complexity; 2) it gives a locally-planar representation; 3) it provides a hierarchical ensemble classification; 4) it allows the application of topologically invariant elementary moves.[1,7–9] In addition, let us stress that *any* network can be embedded on a surface, therefore this approach is completely general and the genus of the embedding surface acts as a constraint on the complexity of the network. Indeed, Ringel and Youngs have shown that an embedding of K_n is always possible in an orientable surface S_g of genus[10]

$$g \geq g^* = \lceil \frac{(n-3)(n-4)}{12} \rceil \tag{1}$$

(for $n \geq 3$ and with $\lceil x \rceil$ the ceiling function which returns the smallest integer number $\geq x$). Therefore, providing a sufficiently high genus $g \geq g^*$, we can always generate embeddings that contain the whole information present in the complete graph. Such a reduction to 2D cannot be in general implemented simply on the plane. Indeed, the complexity of the surface will increase with the complexity of the graph and multi-handled hyperbolic surfaces must be used (for $n > 7$ one has $g^* > 1$, and hyperbolic surfaces are needed to embed the complete graph).

The embedding of the complete graph on S_g is locally planar but this local simplification has been achieved at the expenses of an high complexity in the global surface (g^* scales with n^2). A reduction to simpler surfaces with lower genus is therefore necessary. This can be done only by removing some of the $n(n-1)/2$ links of K_n transforming the complete graph into a less connected network. For instance, the opposite extreme to K_n is the spanning tree which is the connected graph with minimum number of links ($n-1$) and it can be embedded on a surface of genus $g = 0$ (the sphere). Let us now discuss an algorithm[1] which allows to generate networks embedded on surfaces with arbitrary genus and we analyze in details the case $g = 0^2$.

Starting from a correlation coefficient matrix, we construct a graph G embedded on a surface S_g of genus g by means of the following procedure: from a set of disconnected elements, by following an ordered list of pair of elements sorted in decreasing order of the correlation coefficient between two elements, we connect two elements if and only if the resulting network can still be embedded on S_g. Elsewhere the two elements are left disconnected and the subsequent pair in the list is considered. It is important to stress out that this construction is analogous to the procedure which generates the MST. Indeed, the only difference between the two procedures is that in the case of MST a link is inserted if and only if the network after such connection is a forest or a tree.

It has been proved in Tumminello et al.[2] that *at any step of construction of the MST and graph G of genus g if two elements are connected via at least one path in one of the considered graphs then they are connected also in the other one.* This fact implies that the MST is always contained in G. Moreover, this also implies that the formation of connected clusters of nodes during the construction of a graph G coincides with the formation of the same clusters in the MST at the same stage of the construction. In other words the hierarchical structure associated to G coincides with the one of the MST *at any stage of the construction.*

The graph G of genus g is the most connected graph for a given embedding S_g and number of nodes n. The graph G is a triangulation of S_g. It has $3(n - 2 + 2g)$ (for $g \leq g^*$) links and the addition of a further link is impossible without edge-crossings or the increase of the surface genus. A special case is the PMFG[2] which corresponds to $g = 0$, i.e. to a planar embedding.[11] This is the simplest embedding and coincides with the one for the MST. In this respect, the MST and the PMFG are the two extreme cases having the same topological complexity but having the minimum ($n-1$) and the maximum ($3n-6$) number of links respectively.
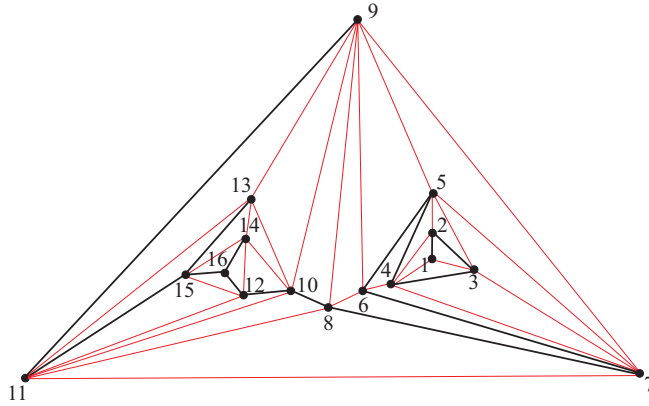
Figure 1. Planar representation of the embedding on S_0 of the PMFG of the 16 Eurodollar interest rates. The labels correspond to the maturity dates expressed in units of 3 months.

The main structural difference between the PMFG and the MST is that the PMFG allows the existence of loops and cliques. A clique of r elements (r-clique) is a complete subgraph that links all r elements. Only cliques of 3 and 4 elements are allowed in the PMFG. Indeed, topological constraints expressed by the Kuratowski's theorem[11] do not allow cliques with a number of elements larger than 4 in a planar graph. Larger cliques can only be present in graphs with genus $g > 0$ and the larger the value of g the larger is the number of elements r of the maximal allowed clique (specifically $r \leq \frac{7+\sqrt{1+48g}}{2}$).[12]

3. STUDY OF THE LOCAL CONNECTIONS AND DISCUSSION

In the previous section we have introduced a *general* method for constructing a network of genus g from the correlation matrix. Here we constraint ourselves to the case $g = 0$, i.e. to the PMFG and we present three examples of PMFG graphs obtained by means of such correlation based procedure. The first example concerns a set of 16 Eurodollars Interest rates.[4, 5] The second example concerns a set of 34 US interest rates[6] and the third concerns a set of 100 US stocks.[2]

3.1. PMFG for 16 Eurodollar interest rates

Let us here start with the analysis of the Eurodollar interest rates. For several economic reasons, interest rates have very similar statistical behaviors and follow similar trends in time. This makes the subject very challenging since one is no more dealing with the statistics of single objects but with the notion of a whole complex set of interacting elements which collectively fluctuate.[5, 6, 13–17] Here we apply our filtering procedure to daily values in the time period $1990 - 1996$ for 16 Eurodollars interest rates with maturity date between 3 to 48 months.[4, 5]

These interest rates are very highly correlated with correlation coefficients values between 0.46 and 0.98 with 76% of the coefficients larger than 0.8. Fig. 1 reports a drawing of the PMFG network on the Euclidean plane. The labels represent the maturity dates (in units of 3 months). In this figure, the thicker lines indicate links belonging to both the MST associated with this set of data and the PMFG. We have therefore constructed a graph which has a larger amount of links respect to the ones contained in the MST. This figure gives an immediate graphical proof that the PMFG network is planar ($g = 0$): indeed it can be drawn on the plane without edge-crossings. In this PMFG we count 38 cliques of 3 elements (triangles). 28 of such triangles lie on the planar surface whereas the other 10 are 'collar rings'. The total number of possible combinations of n elements in groups of three (i.e. the total number of cliques of 3 elements in the complete graph) is $\binom{n}{3} = 560$. This number is much larger than the number of 3-cliques in PMFG (38). Similarly, the number of cliques of 4 elements (tetrahedra) is 10, a number which is much smaller than the number of cliques of 4 elements present in the fully connected graph $\binom{n}{4} = 1820$. Indeed, the planar embedding reduces the interconnectivity of the network simplifying the resulting graph. In Table 1 all the cliques of 4 elements are reported. In this table are
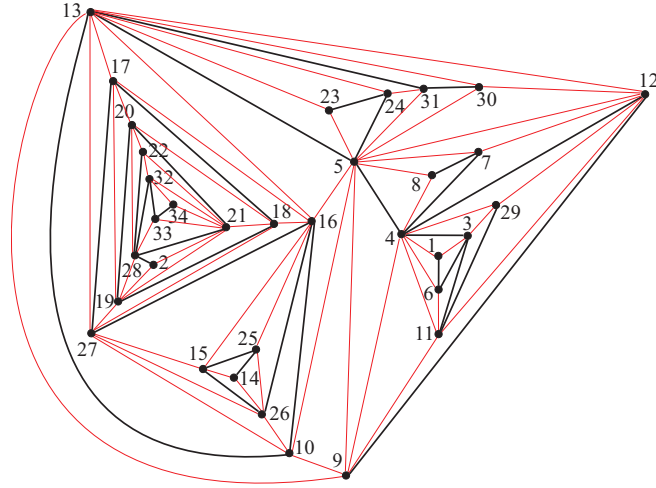
Figure 2. Planar representation of the embedding on S_0 of the PMFG of the correlation structure for the 34 US Interest rates. The labels represent different interest rates described in the Ref..[6]

Table 1. The entire set of 4-cliques in the Eurodollar interest rates. In bold are reported the labels corresponding to maturity dates of 1, 2, 3 and 4 years (labels 4, 8, 12, and 16 respectively.)

		Labels		$\langle c_{i,j} \rangle$	Δ	σ
12	14	15	**16**	0.977	0.017	0.006
9	10	11	13	0.972	0.016	0.006
8	9	10	11	0.969	0.021	0.008
7	**8**	9	11	0.965	0.021	0.007
6	7	**8**	9	0.964	0.026	0.009
5	6	7	9	0.959	0.029	0.012
4	5	6	7	0.958	0.044	0.016
3	**4**	5	7	0.94	0.06	0.02
2	3	**4**	5	0.91	0.13	0.05
1	2	3	**4**	0.83	0.30	0.11

also reported the average correlation coefficients $\langle c_{i,j} \rangle$ inside each clique, the difference Δ between the maximum and minimum correlation coefficient and the standard deviation σ.

## 3.2. PMFG for 34 US interest rates

The second example regards the analysis of correlation coefficients among 34 different weekly interest rates recorded during a time period of 16 years between 1982 and 1997 and stored in the Statistical Release database of the Federal Reserve.[6, 18] For such a data-set we construct both the MST and the PMFG reported in Fig. 2. In this case we observe 94 cliques of 3 elements and 31 cliques of 4 elements. Note that, also in this case, such numbers are much smaller than the number of all possible cliques of 3- and 4-elements in K_{34} which are respectively $\binom{n}{3} = 5984$ and $\binom{n}{4} = 46376$. The complete list of 4 cliques together with $\langle c_{i,j} \rangle$, Δ and σ are listed in Table 2.

3.3. PMFG for 100 US stocks

Let us now show the results of the application of the same method to a much more complex system with a larger number of elements and a larger heterogeneity; a system of 100 stock traded in the New York equity market. We consider pair correlation between daily returns of this set of 100 stocks traded in the time period between 1/1995

**Table 2.** All the4-cliques for the 34 US Interest rates.

|		Labels		$\langle c_{i,j} \rangle$	Δ	σ
22	28	21	20	0.967	0.045	0.016
21	28	20	19	0.97	0.06	0.02
20	21	19	18	0.97	0.05	0.02
9	12	5	4	0.962	0.032	0.013
18	27	17	16	0.96	0.07	0.03
19	20	18	17	0.96	0.06	0.02
26	27	16	15	0.96	0.07	0.03
10	13	9	5	0.957	0.040	0.018
12	13	9	5	0.955	0.040	0.017
30	31	13	5	0.951	0.044	0.017
19	27	18	17	0.95	0.09	0.03
13	30	12	5	0.949	0.044	0.016
25	26	15	14	0.93	0.10	0.05
13	16	10	5	0.93	0.11	0.05
26	27	16	10	0.93	0.11	0.04
17	27	16	13	0.93	0.13	0.05
16	27	13	10	0.93	0.11	0.05
11	12	9	4	0.93	0.10	0.05
7	12	5	4	0.92	0.10	0.04
11	29	4	3	0.92	0.07	0.03
28	32	22	21	0.92	0.11	0.06
6	11	4	3	0.92	0.09	0.03
25	26	16	15	0.92	0.17	0.06
12	29	11	4	0.91	0.10	0.04
7	8	5	4	0.90	0.10	0.04
24	31	13	5	0.87	0.21	0.10
32	33	28	21	0.87	0.17	0.06
21	28	19	2	0.84	0.26	0.13
23	24	13	5	0.80	0.30	0.12
33	34	32	21	0.76	0.23	0.10
4	6	3	1	0.7	0.5	0.2
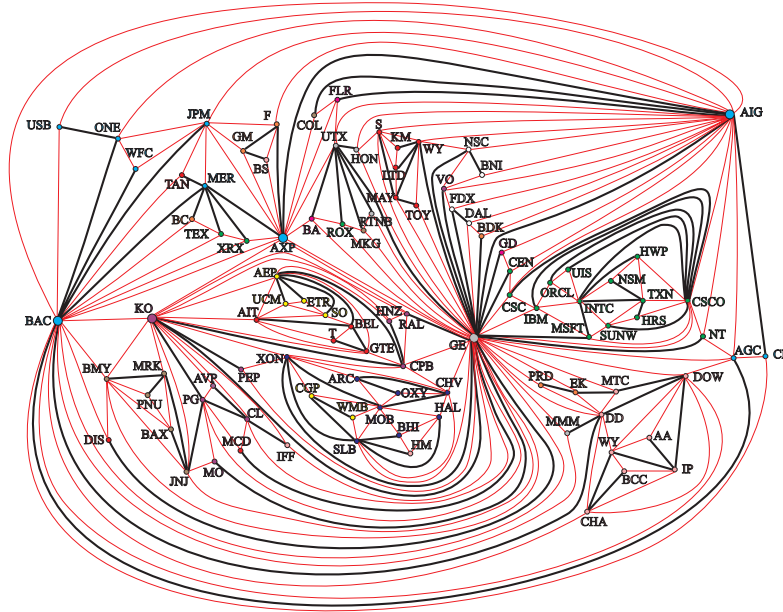
Figure 3. Planar Maximally Filtered Graph obtained from the fully connected graph associated with the correlation coefficient matrix of 100 most capitalized stocks traded in the USA equity markets during the time period from 1995 to 1998.[2] Cross correlation is computed by using daily returns of stocks. Stocks are indicated with their tick symbols.[2] The graph is topologically planar and it can be drawn on the plane without edge-crossings. The thicker lines are belonging to the associated minimum spanning tree. It should be noted that link lengths are not drawn by taking into account the value of the correlation between vertices.

and 12/1998.[19] The planar projection of the PMFG for this system is shown in Fig. 3. As before, in the figure the various elements are connected through links with thicker lines indicating links belonging to both the MST and the PMFG. 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 fully connected graph which are $\binom{n}{4} \cong 3.92 \cdot 10^6$.

The construction algorithm and the topological constraint of the PMFG force each element to participate to at least a clique of 3 elements. The PMFG is a topological triangulation of the sphere. Therefore the *triangular rings* (3-cliques) are the elementary building blocks of this network and all the possible PMFG networks associated to the correlations between a set of n elements can be generated by exploring the ensemble of planar triangulations with n nodes. In a triangulation of the topological sphere the number of triangles on the surface is $2n - 4$. However, in the two systems considered (Eurodollars and US interest rates) as well as in the 100 stock portfolio the number of 3-cliques is systematically larger than the number of triangles on the surface indicating therefore that there are several triangular rings which are collar rings and do not lie on the surface. By observing in details Figs. 4 and 5, where a three-dimensional representation of the PMFG network with the distance between nodes $d_{i,j} = \sqrt{2(1 - c_{i,j})}$ is shown, we note that these internal rings belong to tetrahedra which pack together some 3-clique by sharing a triangular ring which then becomes a collar ring. The basic structures in these graphs 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. If such 4 elements cliques are the 'emerging building blocks' of the PMFG, then there must be strong relations between their properties and the ones of the system from which they have been generated. One can verify that indeed the network for the 34 US interest rates is entirely made by a set of tetrahedra packed together by sharing a triangular face. Such a packing leads to a structure made of $3n - 8$ cliques of 3 elements and $n - 3$ cliques of 4 elements. This is exactly what we find in the case of the 34 US interest rates. Interestingly, the planar network associated with the 16 Eurodollar interest rates does not follow the above scheme having 38 cliques of 3 elements (instead of
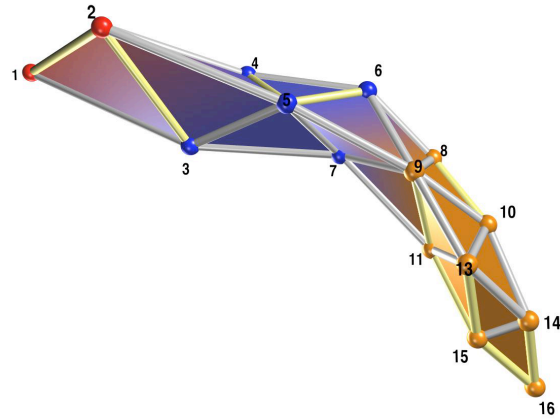
Figure 4. Three dimensional representation of the embedding on S_0 of the PMFG of the correlation structure for the 16 Eurodollar interest rates. Each edge-length corresponds to the metric distance $d_{i,j} = \sqrt{2(1 - c_{i,j})}$. The labels correspond to the maturity dates expressed in units of 3 months.
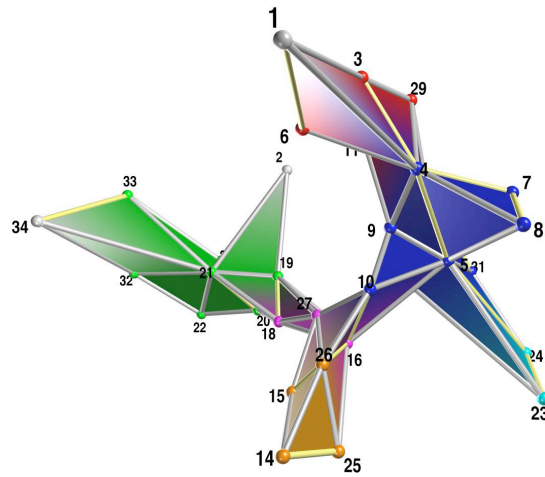
Figure 5. Three dimensional representation of the embedding on S_0 of the PMFG of the correlation structure for the 34 US Interest rates. Each edge-length corresponds to the metric distance $d_{i,j} = \sqrt{2(1 - c_{i,j})}$. The labels represent different interest rates described in the Ref..[6]
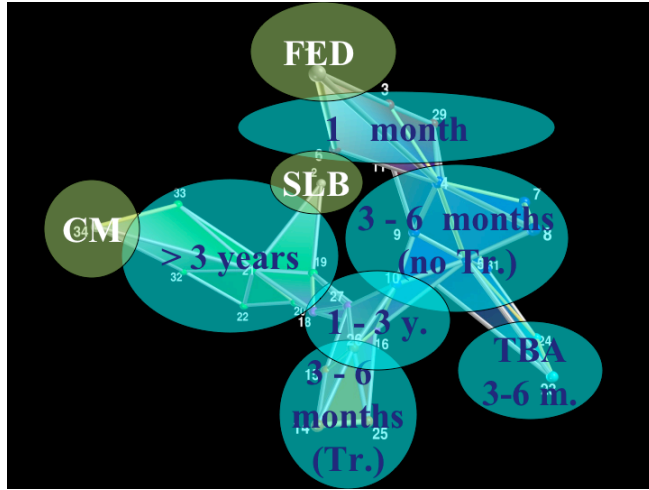
Figure 6. Cluster structure in the PMFG for the 34 US interest rates. The structure can be seen as an aggregate of 4-cliques which reveal the hierarchical organization of the underlying system.

Table 3. Strongest correlated intra-sector 4-cliques. The labels are the economic sectors: Energy (E), Basic Materials (B), Financial (F), Technology (T), Healthcare (H), Consumer non cyclical (C), Services (S) and Utilities (U).

Sec.	Stock 1	Stock 2	Stock 3	Stock 4	Average Correlation
E	ARC	CHV	MOB	XON	0.628
B	BCC	CHA	IP	WY	0.592
F	AXP	BAC	JPM	MER	0.589
T	CSCO	INTC	MSFT	SUNW	0.537
H	BAX	BMY	JNJ	MRK	0.465
C	AVP	CL	KO	PG	0.462
S	AIT	BEL	GTE	T	0.422
U	AEP	ETR	SO	UCM	0.398

$3n - 8 = 40$) and 10 cliques of 4 elements (instead of $n - 3 = 13$). By analyzing the network in detail one can see that such a discrepancy comes from the region where interest rates with large maturity dates of 30-48 months gather together. However, it is clear from Fig. 4 that, apart from this region, the basic structure is also in this case identifiable with a tetrahedral packing.

A further analysis of the clique structure for the case of the Eurodollar interest rates shows that these cliques gather together interest rates forming clusters which are characterized by restricted ranges of maturity dates. From Table 1, one can note that the most correlated cliques are the ones with largest maturity dates and the spread of the correlation coefficients inside each clique increases as the correlation decreases.

A similar differentiation in term of maturity dates arises also from the analysis of the 34 US interest rates. In ref.,[6] it was shown that these data gather together in 6 main clusters and three isolated elements. By comparing the structure of 4-cliques in Table 2 with this cluster gathering we observe that there are 13 cliques composed by elements all belonging to the same cluster, 11 cliques with 3 elements belonging to the same cluster, 5 cliques which mix two elements for one cluster and two from another and finally 2 cliques with elements belonging to three different clusters. Fig.6 shows schematically such a differentiation into different clusters.

The case of the 100 stoks on the US equity market has been discussed in detail in Ref..[2] Let us here briefly recall the main results and observations. An analysis on all the 4-cliques in the PMFG reveals a high degree

of homogeneity with respect to the economic sectors (using the Forbes Magazine classification). The 8 cliques with the largest average correlation for each economic sector with all 4-stocks belonging to the same sector are reported in Table.3. 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 conclusion we have recalled a method to characterize complex systems by means of hyperbolic graphs. We have applied this method to the correlation coefficient matrix in different financial data by analyzing the structure of a network obtained by linking the most correlated elements while constraining the genus of the resulting graph. We have found that in the case $g = 0$ corresponding to the PMFG, the basic structure of such network is formed by packing together cliques of 4 elements which share one or more triangular ring. The study of such 4-cliques in the case of 16 Eurodollar interest rates and 34 interest rates reveals that the network hierarchy spontaneously generates clusters grouping together interest rates with similar maturity dates. Therefore we have shown how this method is efficient in filtering information about the clustering of the system and its hierarchical structure both on the whole system and within each cluster. In particular in the case of 100 US stocks[2] we have shown that the loops and cliques have important and significant relations with the market structure and properties.

REFERENCES

1. T. Aste, T. Di Matteo, S. T. Hyde, "Complex networks on hyperbolic surfaces," *Physica A* **346**, pp. 20-26, 2005.
2. M. Tumminello, T. Aste, T. Di Matteo and R. N. Mantegna, "A tool for filtering information in complex systems, " PNAS 102 (2005) 10421-10426. (cond-mat/0501335).
3. R. N. Mantegna, "Hierarchical structure in financial markets," *Eur. Phys. J. B* **11**, pp. 193-197, 1999.
4. T. Di Matteo , T. Aste, S. T. Hyde a and S. Ramsden, "Interest rates hierarchical structure", *Physica A* **335** pp. 21-33, 2005.
5. T. Di Matteo and T. Aste, "How does the Eurodollars interest rate behave?," *J. Theoret. Appl. Finance* **5**, pp. 122-127, 2002, (cond-mat/0101009).
6. T. Di Matteo, T. Aste, R. N. Mantegna, "An interest rate cluster analysis," *Physica A* **339**, pp. 181-188, 2004, (cond-mat/0401443).
7. H. M. Ohlenbusch, T. Aste, B. Dubertret and N. Rivier, "The topological structure of 2D disordered cellular systems," *Eur. Phys. J. B* **2**, pp. 211-220, 1998.
8. B. Dubertret, T. Aste, H. M. Ohlenbusch and N. Rivier, "Two-dimensional froths and the dynamics of biological tissues," *Phys. Rev. E* **58**, pp. 6368-6378, 1998.
9. T. Aste and D. Sherrington, "Glass transition in self organizing cellular patterns," *J. Phys. A* **32**, pp. 7049-7056, 1999.
10. G. Ringel and J. W. T. Youngs, "Solution of the Heawood Map-Coloring Problem," *Proc. Nat. Acad. Sci. USA* **60**, pp. 438-445, 1968.
11. D. B. West, in *An Introduction to Graph Theory*, Prentice Hall, pp. 247-281, 2001.
12. G. Ringel, in *Map Color Theorem*, Springer, Berlin, Chapter 4, 1974.
13. A. R. Pagan, A. D. Hall, and V. Martin, "Modeling the Term Structure," in *HandBook of Statistics* **14**, G. S. Maddala and C. R. Rao, ed., pp.91-118, 1997.
14. R. Rebonato, *Interest-rate option models*, John Wiley & Sons, New York, 1998.
15. J. P. Bouchaud, N. Sagna, R. Cont, N. EL-Karoui and M. Potters, "Phenomenology of the interest rate curve," *Applied Mathematical Finance* **6**, pp. 209-232, 1999 (cond-mat/9712164).

16. T. Di Matteo, M. Airoldi, E. Scalas, "On pricing interest rate derivatives," *Physica A* **339**, pp. 189-196, 2004 (cond-mat/0401445).

17. T. Alderweireld, J. Nuyts, "Detailed empirical study of the term structure of interest rates. Emergence of power laws and scaling laws," *Physica A* **331**, pp. 602-616, 2004.

18. http://www.federalreserve.gov/releases/h15/data.htm

19. Bonanno, G., Lillo, F., & Mantegna, R., N. (2001) Quantitative Finance **1**, 96-104.