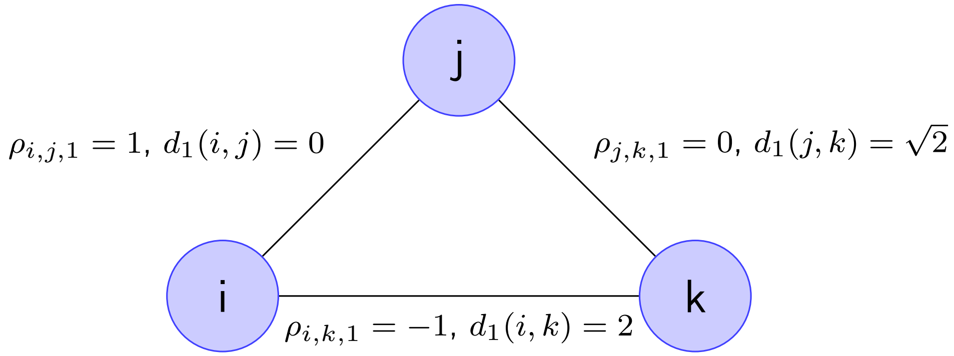
Information Filtering In Networks

Oleg Deev

Network Analysis

- Network is a graph *G(V, E)* describing individual stocks *V* and their relationships *E*
  - For the purposes of our analysis, relationships are set as correlations between returns transformed into distances

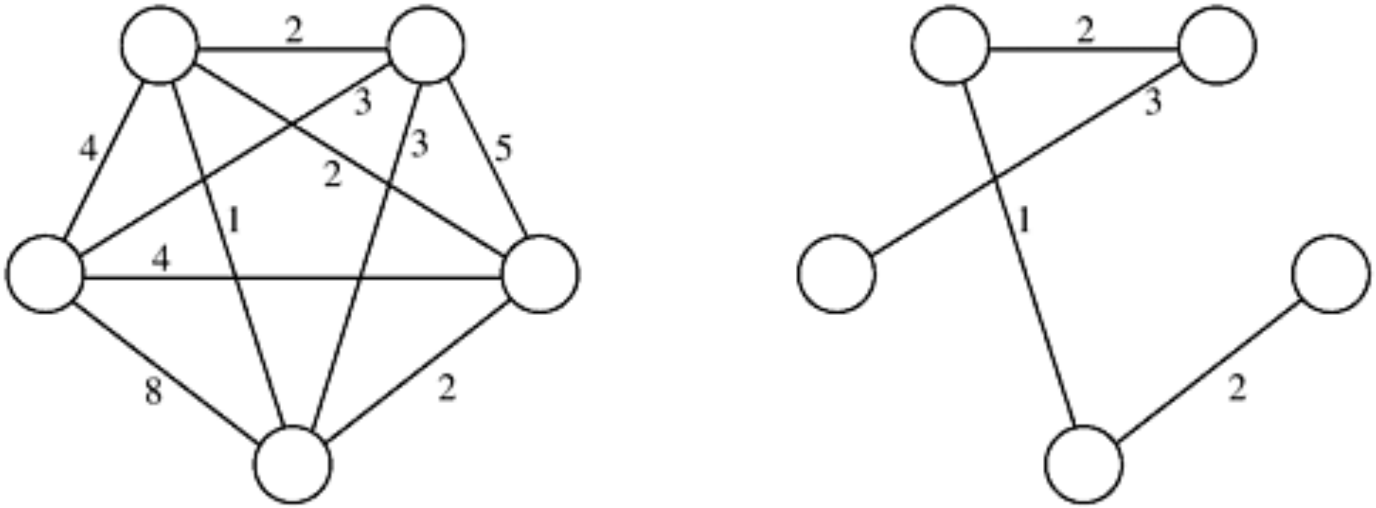
$$d_t(i,j) = \sqrt{2(1-\rho_{i,j,t})}, \qquad i,j \in V, t \in \mathbb{N}$$



- Networks are conceptually simple yet capable of considerable complexity
- **How to subtract the most meaningful information in the network?**

Spanning tree

A spanning tree of an undirected graph *G* is a subgraph of *G* that is a tree containing all the vertices of *G*.

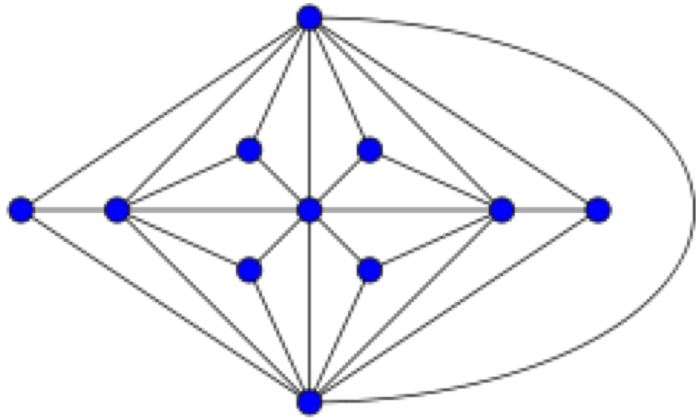


Minimum Spanning Tree (MST)

Step 1. Create an edgeless graph *T = (V, 0)* which vertices correspond with those of *G*.
Step 2. Choose an edge *e* of *G* such that
(i) adding *e* to *T* would not make a cycle in *T* and
(ii) *e* has the minimum weight *w(e)* of all the edges remaining in *G* that fulfill the previous condition.
Step 3. Add the chosen edge *e* to graph *T*.
Step 4. If *T* spans *G*, procedure is terminated; otherwise, the procedure is repeated from Step 2.

Planar graph

- Planar graph is a graph that can be drawn in such a way that no edges cross each other.
- A simple graph is called maximal planar if it is planar but adding any edge (on the given vertex set) would destroy that property.
- All faces (including the outer one) are then bounded by three edges – triangulated graph.



Triangulated Maximally Filtered Graph (TMFG)

uses a structural constraint that limits the number of zero-order correlations included in the network (*3n - 6*; where *n* is the number of variables).

1. Identification of four variables which have the largest sum of correlations to all other variables
2. Iteratively addition of each variable with the largest sum of three correlations to nodes already in the network until all variables have been added to the network

MUNI

Problems with graph-theoretical algorithms

What exactly does the subgraph represent? Take the MST:

- number of edges is fixed
- we do not keep the most important edges
- why insist on connectedness?
- only allows simple structures, no circles/cliques

There is no reason to believe MST represents the structure well

MUNI

Market structure discovery with clique forests

Guido Massara & Tomaso Aste
(UCL Computer Science)

Maximally Filtered Clique Forests (MFCF)

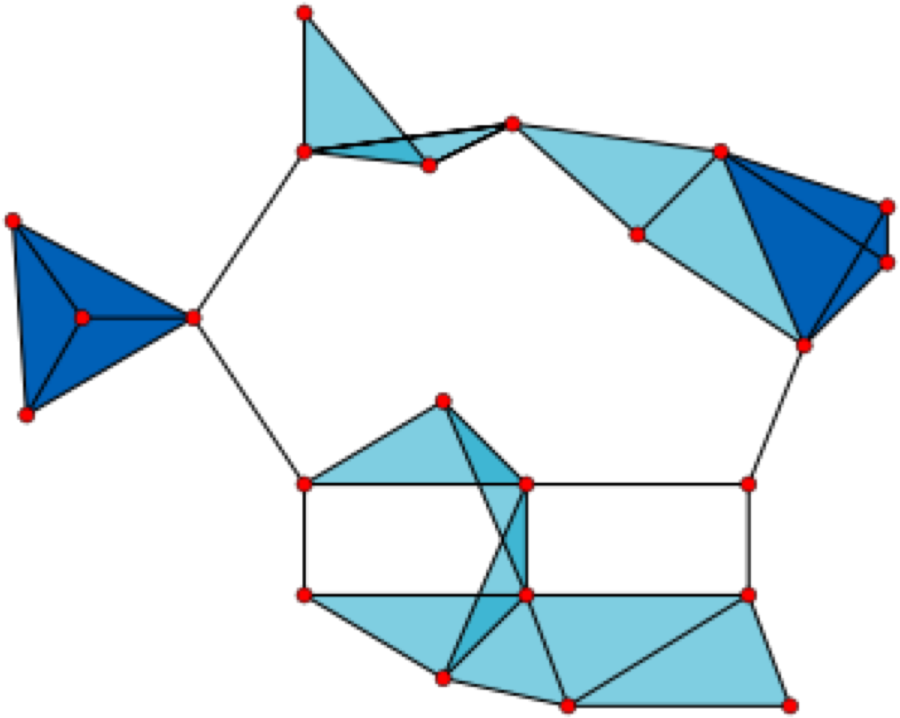
a new and flexible algorithm that allows extracting information from a complex system by imposing topological constraints on the network structure.

The MFCF produces a clique forest. A clique forest is:

- A decomposable graphical model ⇒ probabilistic inference, simulation, what-if analysis
- An information filtering network ⇒ average distance, centrality measures, clustering
- A simplicial complex ⇒ topological data analysis

Clique

- A clique is a subset of vertices $C \subseteq V$ of an undirected graph $G = (V, E)$ such that every two distinct vertices in the clique are adjacent (that is, its induced subgraph is complete).



MUNI

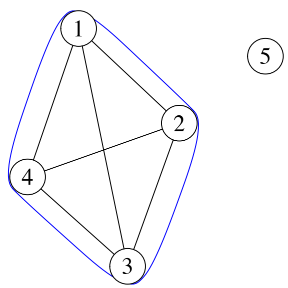
Use of clique forests

Because of their geometric and topological characteristics, clique forests are a highly relevant tools to analyze financial risk management:

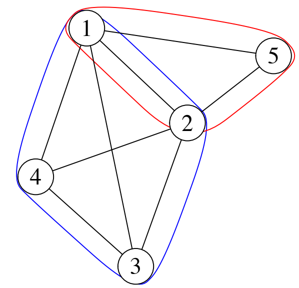
- As a decomposable graphical models they provide a convenient and parsimonious representation of the joint probability distribution of portfolios for asset / risk allocation, risk measurement, stress testing, conditional P/L distribution, scenario analysis, non-normality.
- As information filtering networks they describe the structure of economics and financial networks, individuation of hubs, spillover effects, insight into systemic risk, contagion and macro-prudential regulation.

MFCF construction

The MFCF is based on the repeated application of the clique expansion operator:



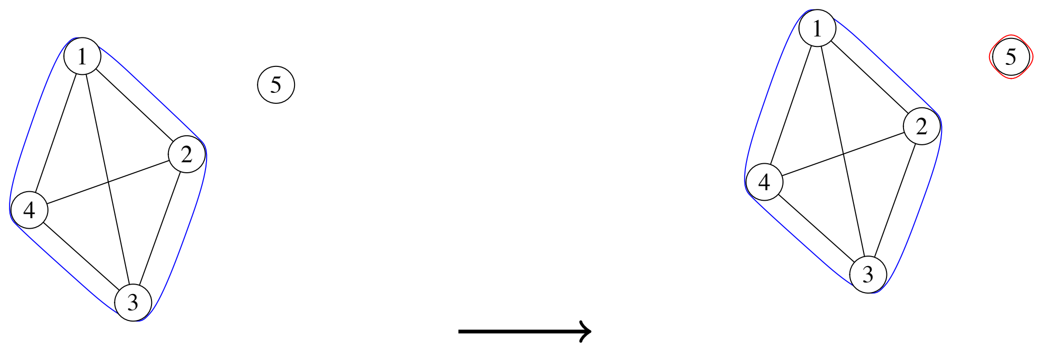
(a) Before clique expansion (general case) $P(X = x \mid G_a) = \phi_{1234}(X_1, X_2, X_3, X_4)\phi_5(X_5)$



(b) After clique expansion (general case), $P(X = x \mid G_b) = \frac{\phi_{1234}(X_1,X_2,X_3,X_4)\phi_{125}(X_1,X_2,X_5)}{\phi_{12}(X_1,X_2)}$ $S = \{1, 2\}$

MFCF construction

The clique expansion operator in case of isolated node:

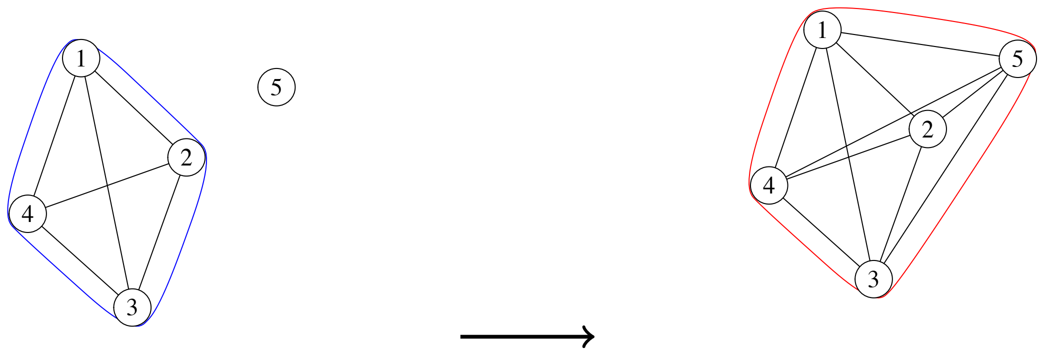


(a) Before clique expansion (isolated vertex case) $P(X = x \mid G_a) = \phi_{1234}(X_1, X_2, X_3, X_4)\phi_5(X_5)$

(b) After clique expansion (isolated vertex case), $P(X = x \mid G_b) = \phi_{1234}(X_1, X_2, X_3, X_4)\phi_5(X_5)$ $S = \emptyset$

MFCF construction

The clique expansion operator in case of full expansion:



(a) Before clique expansion (full expansion)
$P(X = x \mid G_a) = \phi_{1234}(X_1, X_2, X_3, X_4)\phi_5(X_5)$

(b) After clique expansion (full expansion),
$P(X = x \mid G_b) = \phi_{12345}(X_1, X_2, X_3, X_4, X_5)$
$S = \emptyset$

MFCF construction

- The MFCF is driven by a score function (likelihood, sum of weights, AIC, R^2) that can be calculated on the network. Every clique expansion increases the score of the network (gain).
- The MFCF is a greedy algorithm that adds the vertices so that the gain is maximized at every step. Every clique expansion is validated either with a statistical test or by means of cross validation.

Conclusions

- The MFCF is a new tool that produces networks with a very rich structure: probabilistic, geometric / combinatoric, and topological. This allows for a huge range of analyses and applications.
- The algorithm is flexible since it works with a very large range of gain functions: likelihood, R^2, kernel based distances, information measures and entropies, information criteria, regression functions.
- The algorithm is quick and, being local, parallelizable both in discovering the network structure and in using the network for inference.