INTRODUCTION TO GRAPH THEORY AND APPLICATIONS

Andrea Scozzari

University Niccolò Cusano
Dept. of Economics, Psychological and Communication Science.

andrea.scozzari@unicusano.it

MST - Applications

Maximum split Clustering

Consider a set of elements (statistical units) V. For each pair of units i and j in V, a measure of dissimilarity is defined.

$$d(i,j) = \text{dissimilarity between } i \text{ and } j$$

Problem: group the elements of V into k clusters (groups) in such a way that the separation between the clusters is maximized.

How is the separation measured?

MST - Applications

Maximum split Clustering

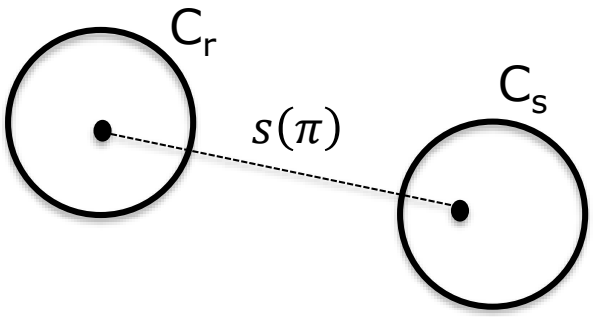
Let P_k be the class of all possible partitions of the elements of V into k clusters.

Let $\pi \in P_k$ be any of these clusters.

The separation associated with π is measured as follows:

$$s(\pi)=\min\{d(i,j) \mid i \in C_r, j \in C_s, C_r \neq C_s\})$$

Where C_r and C_s are any two clusters of π



MST - Applications

Maximum split Clustering

Problem: find $\pi^* \in P_k$ such that:

$$s(\pi^*) = \max \{s(\pi) | \, \pi \in P_k\}$$

Single linkage algorithm:
it is a clustering algorithm: Starting from n clusters each formed by the single elements of V (singletons), at each step the algorithm aggregates the two clusters that have minimum separation.
At each step, the number of clusters decreases by exactly 1.

This is a greedy strategy!

MST - Applications

Maximum split Clustering

Given a complete and weighted graph $G=(V,E)$ that represents the relationship between all pairs of units, let:

$$w(e) = d(i,j)$$

be the dissimilarity between nodes/units.

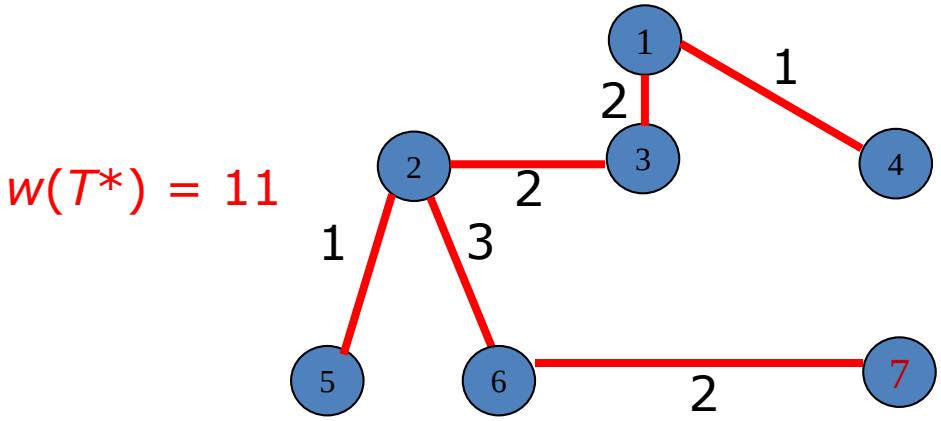
Procedure:

1. Find the MST of G
2. Delete from the MST T^* the k-1 edges of maximum weight

Remark: It can be shown that for any k the procedure finds the same solution $\pi^* \in P_k$ of the Single Linkeage procedure.

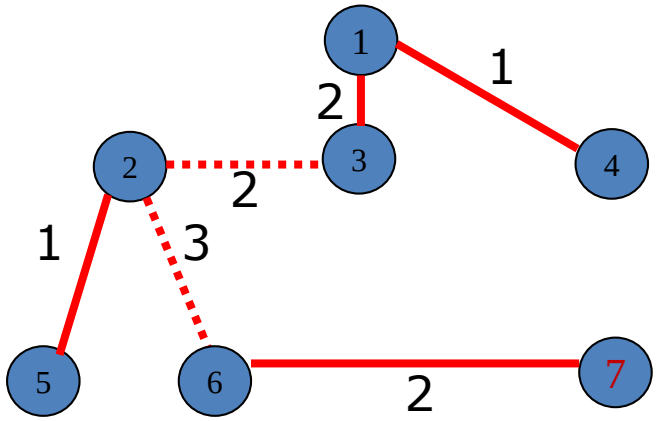
MST - Applications

Let us assume that from a complete graph with n=7 statistical units we have obtained the following MST. Let k=3, i.e. we want to find 3 clusters.



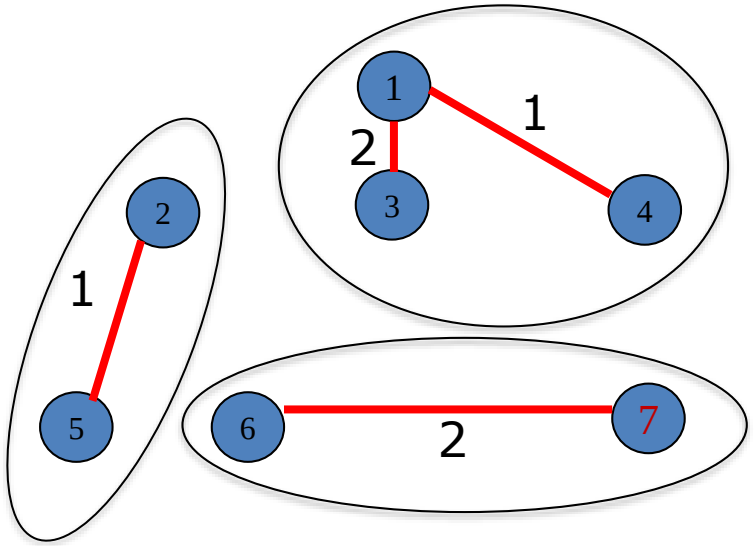
MST - Applications

We have to delete the k-1=2 edges of T^* of maximum weight



MST - Applications

We have to delete the k-1=2 edges of T^* of maximum weight



$s(\pi^*)$, the maximum separation or dissimilarity, between the three groups is equal to the minimum weight of the deleted edges.

$$s(\pi^*)=2$$

MST - Applications

Mantegna, R. N. (1999). Hierarchical structure in financial markets. The European Physical Journal B-Condensed Matter and Complex Systems, 11, 193-197.

Assume that N assets are available. Let R_i i=1,..,N, be the random variables of daily (log-)returns of each asset.

Let $\boldsymbol{P}$ be the Pearson correlation coefficient matrix.

Given an undirected weighted graph G = (V,E,w) without self-loop, where V is the set of vertices/assets and E is the set of edges where i and j are connected by an edge (i,j) and the edges' weights are defined:

$$w_{ij} = \begin{cases} 1 - [\rho(R_i R_j)]^2 & if\ i \neq j \\ 0 & otherwise \end{cases}$$

MST - Applications

Bonanno, G., Lillo, F., & Mantegna, R. N. (2001). High-frequency cross-correlation in a set of stocks. Quantitative Finance, vol.1 96-104.

Assume that N assets are available. Let R_i i=1,..,N, be the random variables of daily (log-)returns of each asset.

Let $\boldsymbol{P}$ be the Pearson correlation coefficient matrix.

Given an undirected weighted graph G = (V,E,w) without self-loop, where V is the set of vertices/assets and E is the set of edges where i and j are connected by an edge (i,j) and the edges' weights are defined:

$$w_{ij} = \begin{cases} \sqrt{2\left(1-\rho\left(R_i R_j\right)\right)} & if\ i \neq j \\ 0 & otherwise \end{cases}$$

MST - Applications

Mantegna, R. N. (1999). Hierarchical structure in financial markets. The European Physical Journal B-Condensed Matter and Complex Systems, 11, 193-197.

The MST of the correlation graph G=(V,E) is attractive because it provides an arrangement of stocks showing the most relevant connections between each node (stock) of V.

Exploiting the MST we can devise a new metric $d_{ij}^{<}$ between any pair of vertices i and j in the MST along with the corresponding matrix $D^{<}$.

$d_{ij}^{<}$ is the *subdominant ultrametric distance* between i and j with respect to the graph G=(V,E):

$$d_{ij}^{<} \geq 0$$

$$d_{ij}^{<} = 0 \Leftrightarrow i = j$$

$$d_{ij}^{<} = d_{ji}^{<}$$

$$d_{ij}^{<} \leq \max\{d_{ik}^{<}, d_{kj}^{<}\}$$

MST - Applications

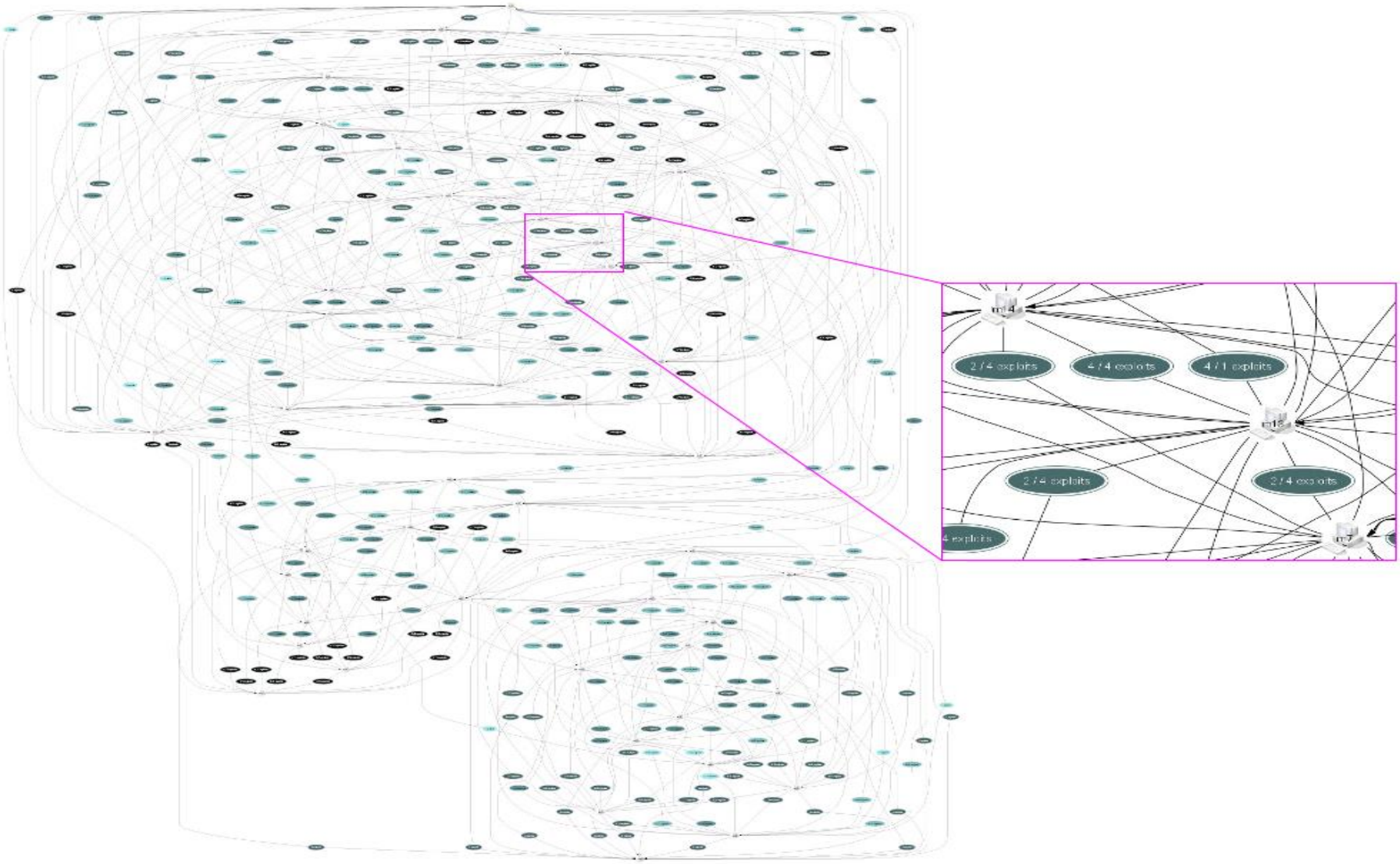
Mantegna, R. N. (1999). Hierarchical structure in financial markets. The European Physical Journal B-Condensed Matter and Complex Systems, 11, 193-197.

The MST of the correlation graph G=(V,E) is attractive because it provides an arrangement of stocks showing the most relevant connections between each node (stock) of V.

With the ultrametric distance, the MST provides, in a direct way, a *hierarchical organization* of the nodes (stocks) of the investigated portfolio.

The Single linkage algorithm, based on the MST, provides a filtering of the stocks, in the sense that, it allows to group assets that are highly correlated, so that, from a diversification viewpoint, one can select just one asset from each group as a representative to form a portfolio with weakly correlated stocks.

CENTRALITY MEASURES



CENTRALITY MEASURES

Given an undirected graph G = (V,E) where E is the set of edges of G and V the set of nodes or vertices of G.

Degree centrality*:* is measured by (simply) considering the degree d_i of a node i, and measures the importance or popularity of a node in the network.

Degree (relative) centrality: is measured by considering the coefficient $\frac{d_i}{n-1}$ of a node i, and measures the relative importance or popularity of a node in the network with respect to the other nodes in the network

CENTRALITY MEASURES

Given an undirected graph G = (V,E) where E is the set of edges of G and V the set of nodes or vertices of G.

Local clustering coefficient: is the fraction of pairs of neighbors of a node i that are connected

$$C_i = \frac{|\{e_{jk} \in E: e_{ij} \in E \cap e_{ik} \in E\}|}{\frac{d_i(d_i - 1)}{2}}$$

where:

d_i = the degree of vertex i

C_i defines the number of triangles in which vertex i participates normalized by the maximum possible number of such triangles.

CENTRALITY MEASURES

Given an undirected graph G = (V,E) where E is the set of edges of G and V the set of nodes or vertices of G.

Global clustering coefficient:

$$C = \frac{1}{n}\sum_{i=1}^{n} C_l(i)$$

A high value of the global index means that any two nodes are connected to other neighboring nodes with high probability.

In social networks, this index describes the number of communities (i.e., groups of nodes) that are closely connected to each other.

CENTRALITY MEASURES: Application

G. Clemente, R. Grassi, A. Hitaj (2019). Asset allocation: New evidence through network approaches. Annals of Operations Research, 299, 61–80.

Assume that N assets are available. Let R_i i=1,..,N, be the random variables of daily (log-)returns of each asset.

Let $\boldsymbol{W}$ be the Pearson correlation coefficient matrix.

Construct an undirected weighted graph G = (V,E,w) without self-loop where V is the set of vertices/assts and E is the set of edges where i and j are connected by an edge (i,j) and the edges' weights are defined:

$$w_{ij} = \begin{cases} \rho(R_i R_j) & if\ i \neq j \\ 0 & otherwise \end{cases}$$

The correlation coefficient of a pair of stocks cannot be used as a distance between the two stocks because it does not fulfill the three axioms that define an Euclidean metric.

CENTRALITY MEASURES: Application

G. Clemente, R. Grassi, A. Hitaj (2019). Asset allocation: New evidence through network approaches. Annals of Operations Research, 299, 61–80.

Consider the new weighting function

$$d_{ij} = \begin{cases} 1 - [\rho(R_i R_j)]^2 & if\ i \neq j \\ 0 & otherwise \end{cases}$$

In the new graph G=(V,E,d) for each node compute the clustering coefficients C_i and define the N-square matrix C, whose elements are:

$$c_{ij} = \begin{cases} C_i C_j & if\ i \neq j \\ 1 & otherwise \end{cases}$$

The matrix C can be interpreted as an interconnectedness matrix.

CENTRALITY MEASURES: Application

G. Clemente, R. Grassi, A. Hitaj (2019). Asset allocation: New evidence through network approaches. Annals of Operations Research, 299, 61–80.

The portfolio selection problem is:

$$\begin{cases} \min x^T H x \\ e^T x = 1 \\ 0 \leq x \leq 1 \end{cases}$$

where:

$H = \Delta^T C \Delta$

$\Delta = \dfrac{\sigma_i}{\sqrt{\sum_{i=1}^{N} \sigma_i^2}} \ (diagonal\ matrix)$

With C and H positive semidefinite.

CENTRALITY MEASURES

Given an undirected graph G = (V,E) where E is the set of edges of G and V the set of nodes or vertices of G, assume G is weighted.
G = (V,E,w).

In such graphs a weight assigned to each edge quantifies, traffic flows passing through the edges, (i.e., air traffic, Internet), strengths of social ties, correlations between stock returns, trade volumes between countries and so forth.

Strength of vertex i $\quad s_i = \sum_{j \in N(i)} w_{ij}$

$N(i)$ =neighbourhood of node i, that is:

$$N(i) = \{j : (i,j) \in E\}$$

CENTRALITY MEASURES

A. Barrat, M. Barth´elemy, R. Pastor-Satorras, and A. Vespignani, Proc. Natl. Acad. Sci. (USA) 101, 3747 (2004)

$$\tilde{C}_B^i = \frac{1}{s_i(d_i - 1)} \sum_{j,k \in N(i)} \frac{w_{ij} + w_{ik}}{2} a_{ij} a_{ik} a_{jk}$$

$a_{ij} = 1$ if edge (i, j) exists, and 0 othewise

J.-P. Onnela, J. Saram¨aki, J. Kert´esz, and K. Kaski, Phys. Rev. E 71, 065103 (2005)

$$\tilde{C}_O^i = \frac{1}{d_i(d_i - 1)} \sum_{j,k \in N(i)} \frac{\left(\sqrt[3]{w_{ij} w_{ik} w_{jk}}\right)}{\max(w)} a_{ij} a_{ik} a_{jk}$$

CENTRALITY MEASURES

B. Zhang and S. Horvath, Stat. App. Genet. Mol. Biol. 4, 17 (2005)

$$\tilde{C}_Z^i = \frac{\sum_{j,k\in N(i)}\left(\widehat{w_{ij}}\widehat{w_{ik}}\widehat{w_{jk}}\right)(a_{ij}a_{ik}a_{jk})}{\sum_{j,k\in N(i)|j\neq k}\left(\widehat{w_{ij}}\widehat{w_{ik}}\right)(a_{ij}a_{ik})}$$

The weights $\widehat{w_{ij}}$ means that they have been normalized by $\max(w)$

P. Holme, S.M. Park, B.J. Kim, and C.R. Edling, Physica A 373, 821 (2007)

$$\tilde{C}_H^i = \frac{\sum_{j,k\in N(i)}\left(\widehat{w_{ij}}\widehat{w_{ik}}\widehat{w_{jk}}\right)(a_{ij}a_{ik}a_{jk})}{\sum_{j,k\in N(i)}\left(\widehat{w_{ij}}\widehat{w_{ik}}\right)(a_{ij}a_{ik})}$$

CENTRALITY MEASURES

Saramäki, J., Kivelä, M., Onnela, J. P., Kaski, K., & Kertesz, J. (2007). Generalizations of the clustering coefficient to weighted complex networks. Physical Review E, 75(2), 027105.

Coeff.	Motivation
$\tilde{C}_B$	Reflects how much of vertex strength is associated with adjacent triangle edges
$\tilde{C}_O$	Reflects how large triangle weights are compared to network maximum
$\tilde{C}_Z$	Purely weight-based; insensitive to additive noise which may result in appearance of "false positive" edges with small weights
$\tilde{C}_H$	Similar to $\tilde{C}_Z$, based only on edge weights

Feature	$\tilde{C}_B$	$\tilde{C}_O$	$\tilde{C}_Z$	$\tilde{C}_H$
1) $\tilde{C} = C$ when weights become binary	X	X	X	
2) $\tilde{C} \in [0, 1]$	X	X	X	
3) Uses global $max(w)$ in normalization		X	X	X
4) Takes into account weights of all edges in triangles		X		X
5) Invariant to weight permutation for one triangle		X		
6) Takes into account weights of edges not participating in any triangle	X		X	X

CENTRALITY MEASURES

Given an undirected graph G = (V,E) with |E|=m and |V|=n the set of nodes or vertices of G.

The assortative mixing index of G provides information about the connection structure of a network.

A network G shows an assortative mixing if in its connection structure there is a prevalence of high-degree nodes attached to other high-degree nodes.

On the contrary, a network G shows a disassortative mixing if high-degree nodes tend to attach to low-degree ones.

CENTRALITY MEASURES

Given an undirected graph G = (V,E) with |E|=m and |V|=n the set of nodes or vertices of G.

The assortative mixing index of G is computed over the edges of a graph as the Pearson correlation coefficient of the degrees of the nodes at either ends of an edge.

$$r = \frac{m^{-1}\sum_{(i,j)\in E} k_i k_j - \left[m^{-1}\sum_{(i,j)\in E} \frac{1}{2}(k_i + k_j)\right]^2}{m^{-1}\sum_{(i,j)\in E} \frac{1}{2}(k_i^2 + k_j^2) - \left[m^{-1}\sum_{(i,j)\in E} \frac{1}{2}(k_i + k_j)\right]^2}$$

where:

k_i = the excess degree of vertex i computed as: d_i-1

d_i = the degree of vertex i

It measures a preference for a network's nodes to attach to other nodes that are similar in some way.

CENTRALITY MEASURES: Application

Ricca, F., & Scozzari, A. (2024). Portfolio optimization through a network approach: Network assortative mixing and portfolio diversification. European Journal of Operational Research, 312(2), 700-717.

Given an undirected graph G = (V,E) with |E|=m and |V|=n the set of nodes or vertices of G.

The assortative mixing index of G varies in [−1, 1] and it is equal to 1 for a graph with a perfect assortative mixing, equal to −1 when the graph is perfectly disassortative, and it is 0 for networks which do not present any clear tendency among the previous two.

The local assortativity coefficient of a node i is formulated as follows:

$$r_i = \frac{d_i k_i (\bar{k}_{N(i)} - \mu_{q(k)})}{2m\sigma^2_{q(k)}}$$

$\mu_{q(k)}$ and $\sigma^2_{q(k)}$ are the mean and the variance of the nodes' excess degree
$q(k)$ is the empirical probability distribution of the nodes' excess degree.
$\bar{k}_{N(i)}$ is the average excess degree of the nodes in the neighborhood of i.

CENTRALITY MEASURES

Ricca, F., & Scozzari, A. (2024). Portfolio optimization through a network approach: Network assortative mixing and portfolio diversification. European Journal of Operational Research, 312(2), 700-717.

We proposed a portfolio selection model with objective function:

$$\max \sum_{i=1}^{N} \mu_i x_i - \sum_{i=1}^{N} r_i x_i$$
$$\boldsymbol{x} \in \boldsymbol{\Omega}$$

The search for the optimal solution is guided towards a high portfolio expected return, and, simultaneously, to a selection of disassortative assets.

CENTRALITY MEASURES

Ricca, F., & Scozzari, A. (2024). Portfolio optimization through a network approach: Network assortative mixing and portfolio diversification. European Journal of Operational Research, 312(2), 700-717.

Strength of vertex i

$$s_i = \sum_{j \in N(i)} w_{ij}$$

local strength assortativity index which corresponds to the single contribution of each node in the network to the strength assortative mixing value of the graph.

$$Sr_i = \frac{es_i \sum_{j \in N(i)} es_j - d_i \, es_i \, \mu_{q(s)}}{2m\sigma^2_{q(s)}}$$

$\mu_{q(s)}$ and $\sigma^2_{q(s)}$ are the mean and the variance of the nodes' excess strength

$q(s)$ is the empirical probability distribution of the nodes' excess strength.

$\sum_{j \in N(i)} es_i$ is the average excess strength of the nodes in the neighborhood of i.

CENTRALITY MEASURES

Ricca, F., & Scozzari, A. (2024). Portfolio optimization through a network approach: Network assortative mixing and portfolio diversification. European Journal of Operational Research, 312(2), 700-717.

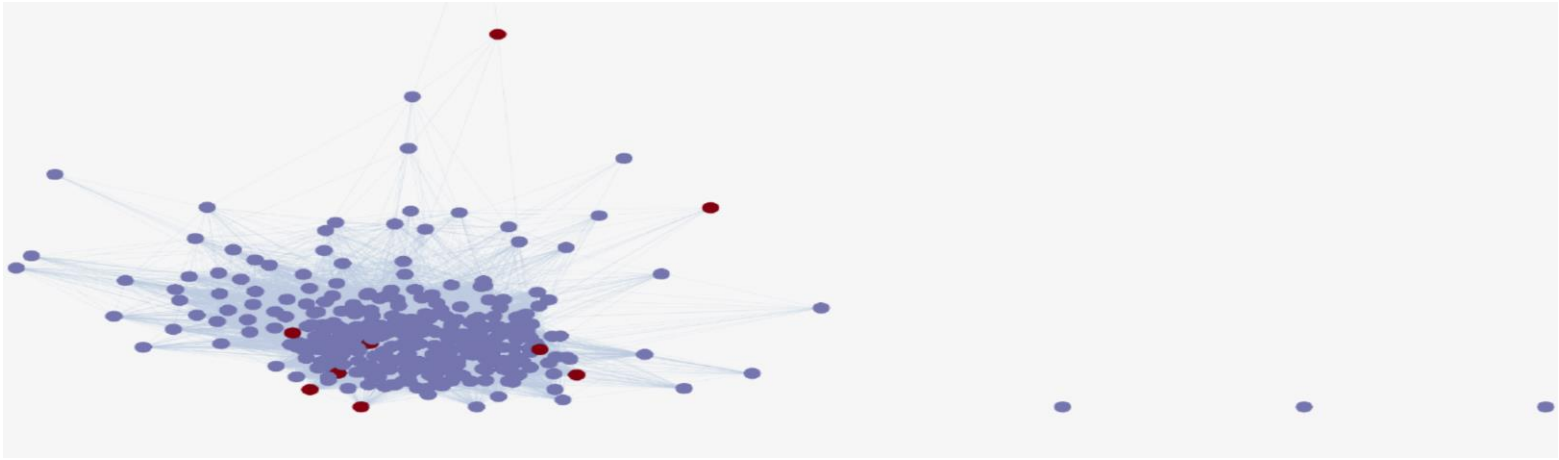
We proposed a portfolio selection model with objective function:

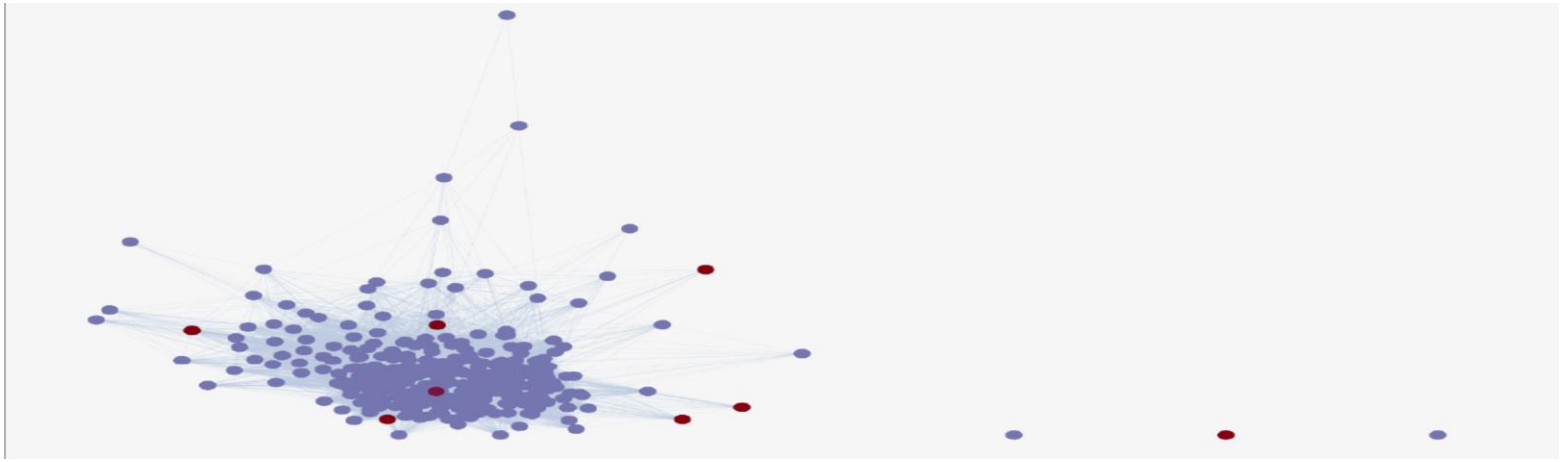
$$\max \sum_{i=1}^{N} \mu_i x_i - \sum_{i=1}^{N} Sr_i x_i$$
$$\boldsymbol{x} \in \boldsymbol{\Omega}$$

The search for the optimal solution is guided towards a high portfolio expected return, and, simultaneously, to a selection of disassortative assets.

CENTRALITY MEASURES

Ricca, F., & Scozzari, A. (2024). Portfolio optimization through a network approach: Network assortative mixing and portfolio diversification. European Journal of Operational Research, 312(2), 700-717.





GRAPH CLASSES: Scale Free Networks

https://www.graphclasses.org/smallgraphs.html#header.

An undirected graph G = (V,E) is Scale-free when its node-degree distribution follows a power-law, at least asymptotically.

$$P_{deg}(k) \propto k^{-\gamma}$$

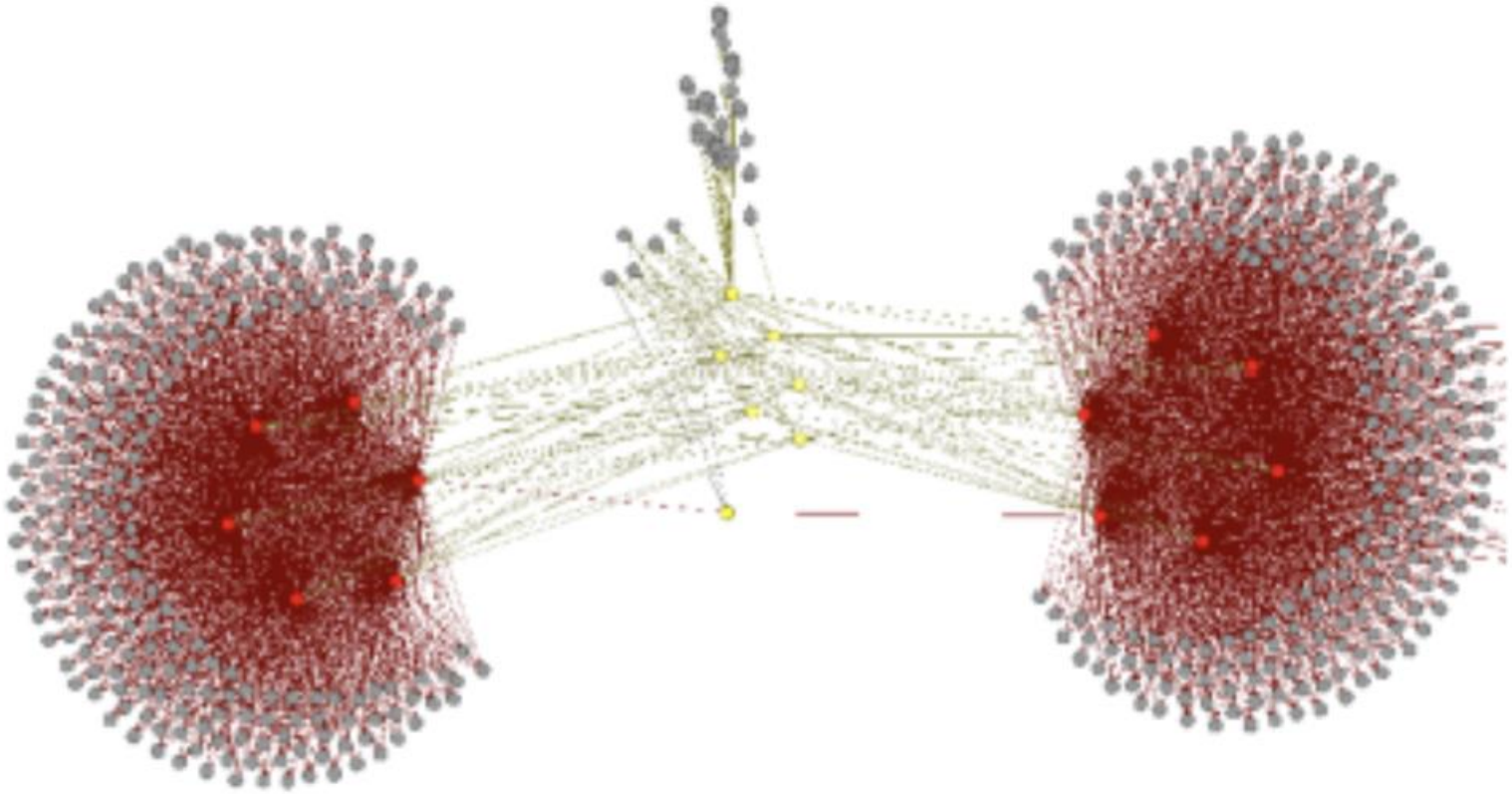
where γ is some exponent.

This form of the probability distribution decays slowly as the degree k increases, increasing the likelihood of finding a node with a very large degree.

Networks with power-law distributions are called scale-free because power laws have the same functional form at all scales.

GRAPH CLASSES: Scale Free Networks

Gulino, A., Ceri, S., Gottlob, G., Sallinger, E., & Bellomarini, L. (2021, April). Distributed company control in company shareholding graphs. In 2021 IEEE 37th International Conference on Data Engineering (ICDE).



GRAPH CLASSES: Small Worlds Networks

https://www.graphclasses.org/smallgraphs.html#header.

An undirected graph G = (V,E) has a small-world property if it has a high (global) clustering coefficient and a small path length.

A high clustering coefficient means that the graph G as a high number of cliques and "near-cliques" resulting in subnetworks comprising edges between all or almost all vertices.

A small path length represents a global reachability property that is the average vertex-to-vertex distances increases only logarithmically with respect to the total number |V| of vertices of G.