Research and development of algorithms using cluster-based interactions of metagenomic data in biomedicine

Camille Champion

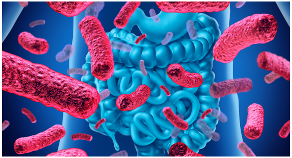
20/09/2021

Institut des Maladies Métaboliques et Cardiovasculaires
INSTITUT de MATHEMATIQUES de TOULOUSE
INRAE

Biological context

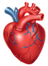
Microbial composition reflects :

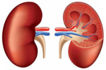
- environment,
- lifestyle,
- metabolism,
- diseases,

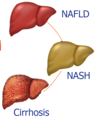


Diseases associated with imbalance microbiota :

- Cardio-vascular diseases,
- Kidney diseases,
- Metabolic diseases.
  - Obesity,
  - Diabetes,
  - Cirrhosis.







Objective

Find biological signatures related to the development of metabolic and cardiovascular diseases

Biological system modelling

A **biological system** with :

- p quantitative variables : $X^1, \dots, X^p$,
- n observations : $X_1^j, \dots, X_n^j,\ j \in [\![1, p]\!]$,

modeled by **undirected graphs** $G(V, E)$ with no self-loops where :

- one vertex=one gene or metagene,
- one edge=one connection between two genes,
- $V = \{1, \dots, p\}$ and E are the vertices and edges set.

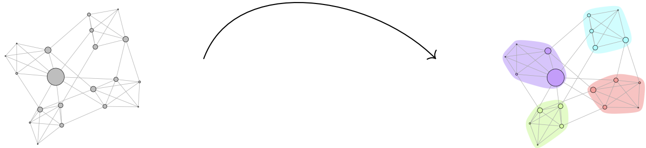
Objective :

- Model the functional relationships between the composing elements of the system,
- Emphasize major interactions,
- Understand the underlying biological processes.

Graph Clustering

Graph Clustering

- From a **graphical** point of view, cluster vertices into groups that are densely connected and share a few links (comparatively) with the other groups,
- From a **biological** point of view, discover groups of genes with similar characteristics to better understand a disease.



Wide range of very popular clustering algorithms based on graph-theory :

- **Partitioning algorithms (*k*-means) :** classify nodes into a predefined number of groups based on a similarity measure (MacQueen, 1967),
- **Spectral clustering algorithms :** use the spectral properties of the graph to recover the graph structure (Luxburg, 2007).

Contributions

1. CORE-clustering algorithms and applications,
   - Algorithms for the detection of representative variables in complex systems,
   - Application to simulated data and a road network.
2. ℓ_1-spectral clustering algorithm and applications,
   - A robust spectral clustering using LASSO regularization,
   - Application to simulated data and kidney cancer.
3. Human liver microbiota modeling strategy at the early onset of fibrosis.

Detection of Representative Variables in Complex Systems with Interpretable Rules Using Core-Clusters

CORE-clustering algorithm

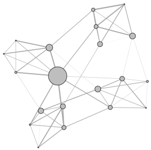
Graph-based representation, issues and objective

A complex system ($n << p$) modeled by an **undirected weighted graph** $G(V, E)$ made of a set V of vertices $(X^1, \ldots, X^p)$ and a set E of edges.

Goal : Detection of interpretable cluster structures in a high dimensional graph

Issues

- Instability due to the high complexity of the system,
- Choice of the granularity level,
- Interpretability of the clusters found.



Key solution : Robust detection of clusters structured around representative variables of the complex system

Coherence in a subset

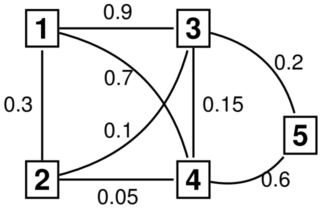
Definition

*A **path** P of a graph G from X^i to X^j of length Λ is a list of indices* $\{d_1, \dots, d_\Lambda\} \subset [\![1, p]\!]$ *such that :* $\begin{cases} X^i = X^{d_1}, \\ X^j = X^{d_\Lambda}. \end{cases}$

Definition

*The **path capacity** c(P) is the minimal weight of the edges through which P passes :*

$$cap(P) = \min_{l=1,\dots,\Lambda-1} w_{d_l, d_{l+1}}. \tag{1}$$



Path : $\{1, 3, 4, 5\}$

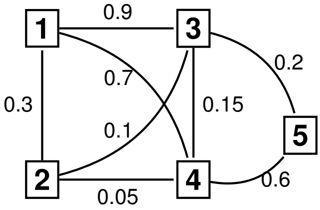
Capacity : 0.15

Coherence in a subset

Definition

*The **coherence** $c(X^i, X^j)$ between X^i and X^j is defined by considering the path P having the maximum capacity among the paths of $\mathbf{P}_{i,j}$:*

$$c(X^i, X^j) = \max_{P \in \mathbf{P}_{i,j}} cap(P). \tag{2}$$



Coherence between nodes : 1 and 5

Coherence : 0.6

Path with maximal capacity : $\{1, 4, 5\}$

Definition

*The **coherence** $\mathbf{c}(S)$ of the **variable subset** S is the minimal coherence between the variables it contains :*

$$\mathbf{c}(S) = \min_{(X^i, X^j) \in S^2} c(X^i, X^j). \tag{3}$$

CORE-Clusters

Definition

- *A* ***CORE-cluster*** *is a variable subset $S \subset X$ respecting the following properties :*
  - *its size is in the range $[\tau, 2\tau - 1]$,*
  - *its coherence is higher than a threshold ξ.*

  $\Big\}$ *τ and ξ are tuning parameters*
- *A* ***representative variable*** *is defined as centred CORE-cluster center..*

Estimation of an optimal set of CORE-clusters $\widehat{\mathbf{S}} = \{\widehat{S^u}\}_{u \in \{1,\ldots,\widehat{U}\}}$:

$$\left(\widehat{\mathbf{S}}, \widehat{U}\right) = \underset{(\mathbf{S},U)}{\arg\max} \sum_{u=1}^{U} \mathbf{c}(S^u) \tag{4}$$

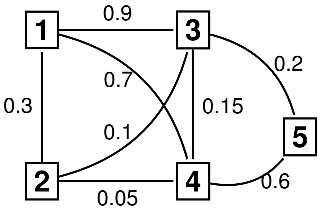
under the two constraints :

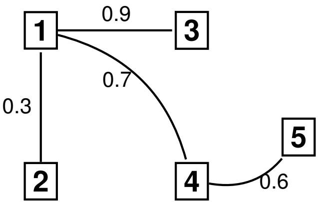
1. CORE-clusters $S^u_{\xi,\tau}$ have a size higher than τ and a coherence $\mathbf{c}(S^u_{\xi,\tau}) > \xi$,
2. No overlap between the clusters, *i.e.* $\forall (u_1, u_2) \in \{1, \ldots, U\}^2, S^{u_1} \cap S^{u_2} = \emptyset$.

Maximum Spanning Tree (Kruskal, 1956)

Definition

- *A spanning tree $G(V,T)$ is a connected subgraph of $G(V,E)$ with* $\begin{cases} \text{no cycle}, \\ T \subset E. \end{cases}$
- *A maximum spanning tree of G is the spanning tree of G having the maximal sum of edge weights*

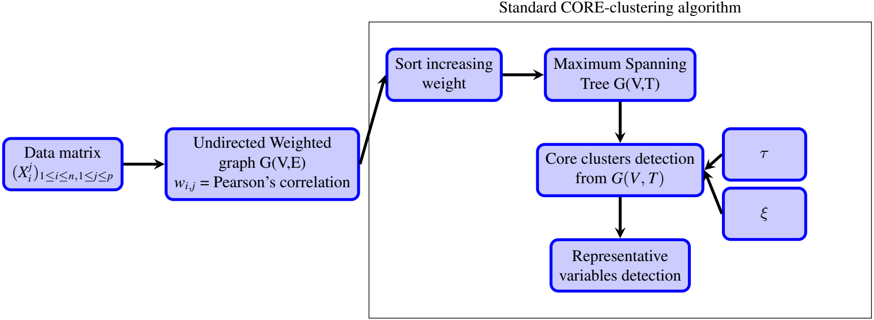




Core-clustering algorithm main steps

Input parameters :

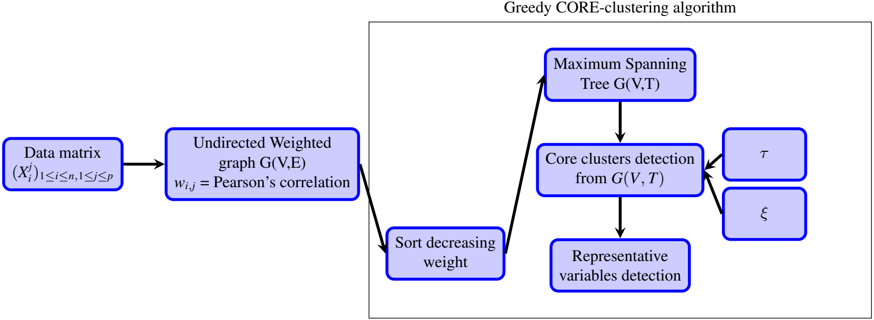
- Minimal dimension of the core-clusters (τ)
- Minimum level of similarity which gathers their variables (ξ)



Core-clustering algorithm main steps

Input parameters :

- Minimal dimension of the core-clusters (τ)
- Minimum level of similarity which gathers their variables (ξ)



Core detection in synthetic data

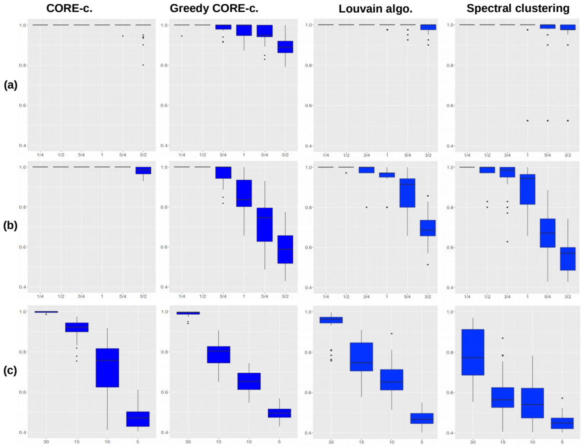


FIGURE – (**a**) Two simulated clusters with noise levels ranging from 0.25 to 1.5. (**b**) Same as (**a**) with five simulated clusters. (**c**) Five clusters simulated using 30, 15, 10 and 5 observations of $[250, 500]$ variables and a noise level of 0.5.

ℓ_1-spectral clustering : a robust spectral clustering using LASSO regularization

ℓ_1-spectral clustering algorithm

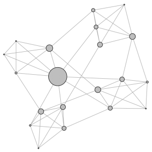
Graph-based representation, issues and objective

A system modeled by an undirected unweighted graph $G(V, E)$ made of a set V of vertices $(X^1, \ldots, X^p)$ and a set E of edges.

Goal : Detection of interpretable cluster structures in a noisy graph

Issues

- Noise sensitivity of spectral clustering algorithm,
- Choice of the number of clusters,
- Interpretability of the clusters found.



Key solution : Detection of cluster structures in a noisy graph using a spectral clustering variant

Adjacency and Laplacian matrices

Definition

- *The **adjacency matrix** A of G is defined as :*

$$\forall (i,j) \in [\![1,p]\!]^2, \;\; A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Definition

- *The **degree** d_i of vertex X^i is the number of edges incident to i*

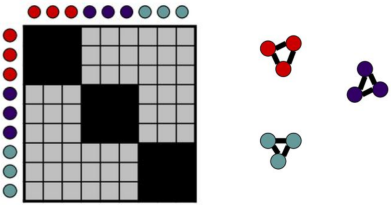
$$d_i = \sum_{j=1}^{p} A_{ij} \;\; \text{and } D \text{ as the associated degree matrix.}$$

- *The **Laplacian matrix** L of G is defined as : $L = D - A$, where D the degree matrix and A the adjacency matrix associated to G.*

Graphs : assumptions

The unknown structure of the graph G to cluster is assumed to be made of k connected components $C_1, \dots, C_k$.

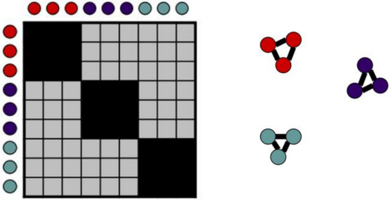
$$A = \begin{bmatrix} \underbrace{\begin{matrix} 0 & 1 & \cdots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 0 \end{matrix}}_{c_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \underbrace{\begin{matrix} 0 & 1 & \cdots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 0 \end{matrix}}_{c_k} \end{bmatrix}$$



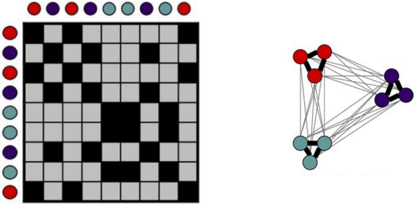
Graphs : assumptions

The unknown structure of the graph G to cluster is assumed to be made of k connected components $C_1, \ldots, C_k$.

$$A = \begin{bmatrix} \underbrace{\begin{matrix} 0 & 1 & \cdots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 0 \end{matrix}}_{c_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \underbrace{\begin{matrix} 0 & 1 & \cdots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 0 \end{matrix}}_{c_k} \end{bmatrix}$$



Perturbed graph : Let $\hat{G}$ be a perturbed version of G, obtained by adding/removing an edge between/inside components of the graph with probabilities $(p_{in}, p_{out}) \in [0, 1]^2$.



Spectral clustering algorithm

Properties of the Laplacian matrix

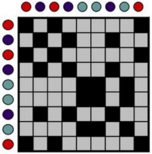
- L is symmetric and positive semi-definite,
- L has p non-negative real-valued eigenvalues $\lambda_1, \ldots, \lambda_p$,
- The smallest eigenvalue of L is 0.

Proposition

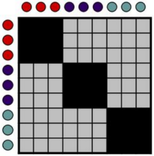
- *The eigenvalue* 0 *of* L *is of multiplicity* k *(number of connected components),*
- *The associated eigenvectors correspond to the indicator vectors* $(1_{C_i})_{1 \leq i \leq p}$ *of the* k *components.*

Goal : Cluster the nodes of a graph $G(V, E)$ into k communities

$$L \xrightarrow[\text{eigenvalue } 0]{\text{SVD}}$$



$$\xrightarrow{k\text{-means}}$$



Advantages, issues and alternatives

Advantages and issues : Spectral clustering on the perturbed version of the graph

- Refinements using the normalized versions of the Laplacian matrix (Symmetric, Random Walk normalized Laplacian matrices,...),
- Powerful computational results,
- Theoretical convergence results,
- High sensitivity and no guarantee of recovering the true components in case of large perturbations.

Alternatives : Development of the ℓ_1-spectral clustering new algorithm

- Laplacian matrix replaced by Adjacency matrix,
- k-means procedure replaced by the selection of relevant eigenvectors, solutions to specific ℓ_1-minimization problems.

Theoretical results I

We denote by

- $\lambda_1, \ldots, \lambda_p$ the p eigenvalues of the adjacency matrix A,
- $v_1, \ldots, v_p$ the associated eigenvectors,
- $\mathcal{V}_k$ the eigenspace generated by the k largest eigenvectors :

$$\mathcal{V}_k = Span(v_{n-k+1}, \ldots, v_p).$$

Proposition

The minimization problem $(\mathcal{P}_0)$

$$\underset{v \in \mathcal{V}_k \setminus \{0\}}{\arg\min} \; \|v\|_0$$

has a unique solution (up to a constant) given by 1_{C_1}.

Theoretical results II

We denote by

- $\lambda_1, \ldots, \lambda_p$ the p eigenvalues of the adjacency matrix A,
- $v_1, \ldots, v_p$ the associated eigenvectors,
- $\mathcal{V}_k$ the eigenspace generated by the k largest eigenvectors :

$$\mathcal{V}_k = Span(v_{n-k+1}, \ldots, v_p).$$

From now on, we assume that we know a node belonging to each component, called **representative element** and denoted by $(i_1, \ldots, i_k)$. Let $\tilde{\mathcal{V}}_k$ be :

$$\tilde{\mathcal{V}}_k := \{v \in \mathcal{V}_k, v_{i_1} = 1\}.$$

Proposition

The minimization problem $(\mathcal{P}_1)$

$$\underset{v \in \tilde{\mathcal{V}}_k}{\arg\min} \; \|v\|_1$$

has a unique solution given by 1_{C_1}.

Theoretical results III

Proposition

Let $U_k := (v_1, ..., v_{p-k})$ the matrix formed by the eigenvectors associated with the $p-k$-smallest eigenvalues. We denote by w^T its first row and W^T the matrix obtained after removing w^T from U_k :

$$U_k := (v_1, ..., v_{p-k}) = \begin{bmatrix} \boxed{w^T} \\ \boxed{W^T} \end{bmatrix} \tag{5}$$

The minimization problem

$$\underset{\substack{v \in \mathbb{R}^{p-1} \\ Wv = -w}}{\arg\min} \; \|v\|_1 \tag{$\tilde{\mathcal{P}}_1$}$$

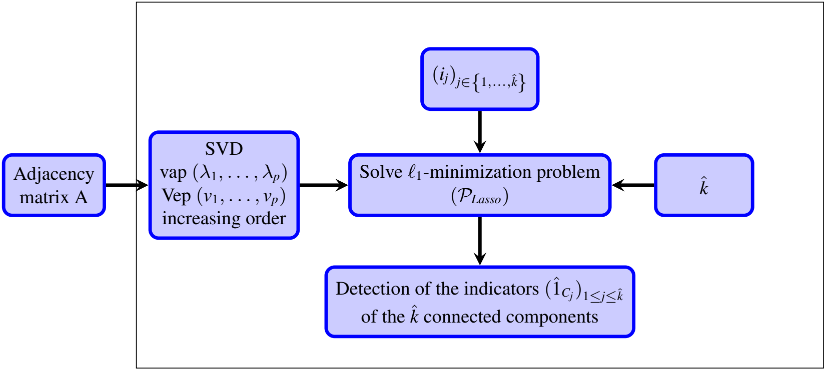
has a unique solution v^* such that $(1, v^*)^T = 1_{C_1}$.

ℓ_1-spectral clustering algorithm main steps

Input parameters :

- Number of clusters $\hat{k}$ to recover,
- $(i_j)_{j \in \{1,\ldots,\hat{k}\}}$ family of representative elements of each cluster found using a betweeness centrality score.

ℓ_1-clustering algorithm



Comparison with state-of-the-art

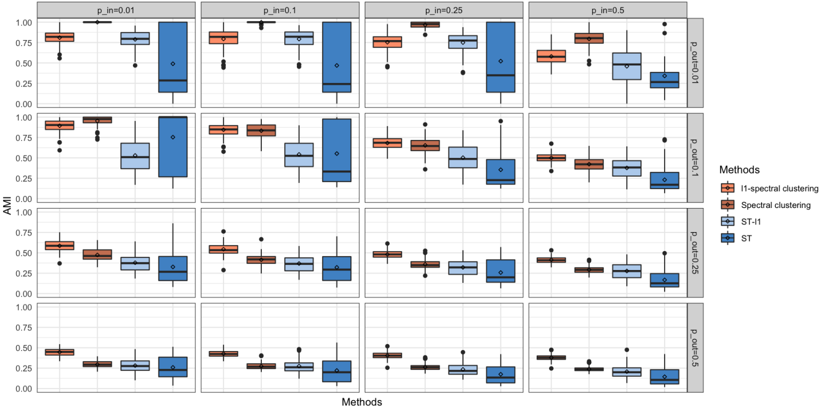


FIGURE – Simulation of 100 versions of the same perturbed graphs with $p = 50$ variables, $k = 10$ components and perturbations p_{in} and p_{out} of removing/introducing an edge from/between components varying from 0.01 to 0.5.

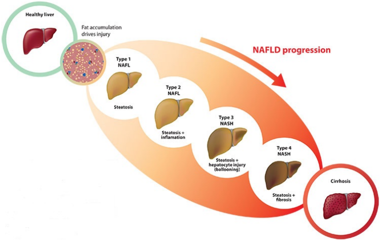
Modeling of liver microbiota at the early onset of human fibrosis

Statistical study of liver fibrosis cohort

Overview

A 82 cohort affected, at various stages, by liver fibrosis :

- F0 : no Fibrosis
- F1 : minor Fibrosis
- F2 : moderate Fibrosis



Liver Fibrosis

Formation of an abnormally large amount of scar tissue in the liver. It occurs when the liver attempts to repair and replace damaged cells.

Goal : Identify the patients' clinical phenotypic profile and the microbial species involved in the early onset of the disease

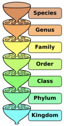
Datasets

Clinical features :

- Hypertension
- Dyslipidemia
- Diastolic
- Systolic
- Diabete
- Blood-glucose
- Age

Metagenomic features :

- OTU table count at different levels
- Taxonomy



Definition (Operational Taxonomic Units)

Cluster of similar sequence variants of the 16S rDNA marker gene sequence (97%).

1. DNA extraction,
2. 16S gene amplification + sequencing of some regions,
3. Partitionning of reads (nucleotide sequences) into OTUs,
4. Taxonomic assignations.

Statistical analysis adapted to metagenomic datasets

- Exploratory analysis (PCA, (Pearson, 1901)),

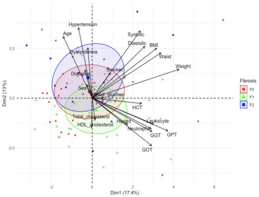
Goal : Identify clinical phenotypic and bacterial profile of fibrotic patients

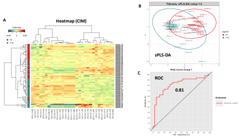
- Discriminant analysis (PLS-DA and variants, (Barker and Rayens, 2003)),

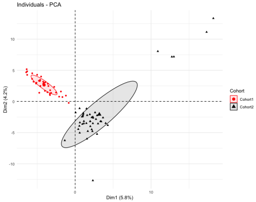
Goal : Detect microbial species and functional metabolic pathways involved in the development of the disease

- Fair exploratory and discriminant analysis (fair PCA, ℓ_1-spectral clustering and fairlet clustering),

Goal : Address the bias effect generated by the population's diversity and explain the total variabilities in the dataset







Conclusion and outlooks

Work already done and under development :

- Development of two graph clustering algorithms to detect highly connected groups of variables :
  - Core-clustering within a high dimensional complex system,
  - ℓ_1-spectral clustering within a noisy graph.
- Statistical analysis of a cohort of liver fibrotic patients to discover biological signatures categorizing patients in the disease :
  - Standard exploratory, discriminant, clustering methods (PCA, PLS-DA),
  - New fair approach based on exporatory and regression techniques,

Perspectives :

- Adaptation and application of graph clustering methods (CORE-clustering and ℓ_1-spectral clustering) to bacterial datasets.

Thanks for your attention !

- P.K. Agarwal, S. Har-Peled, K.R. Varadarajan (2005). Geometric approximation via coresets. Combinatorial and Computational Geometry, MSRI. University Press, 1–3.
- C. Champion, A.C. Brunet, R. Burcelin, J.M. Loubes, L. Risser (2021). Detection of Representative Variables in Complex Systems with Interpretable Rules Using Core-Clusters. Algorithms 14 (2), 66.
- C. Champion, M. Champion, M. Blazère, R. Burcelin, JM. Loubes. l_1-spectral clustering algorithm : a robust spectral clustering using Lasso regularization. Submitted, 2021.
- C. Champion and Al. Human liver microbiota modeling strategy at the early onset of fibrosis. Submitted, 2021.
- J.B. Kruskal (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical Society, 7 : 48–50.
- Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and Computing 17(4), 395—416.
- MacQueen, B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1, 281–297.
- Seidman, S.B. (1983). Network structure and minimum degree. Social Networks 5(3), 269–287.
- Ward, J. (1963). Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 58(301), 236–244.