Data analysis for labeled graphs

Nathalie Villa-Vialaneix
http://www.nathalievilla.org
nathalie.villa@univ-paris1.fr

44èmes JdS Bruxelles - 23 mai 2012

Joint work with **Thibault Laurent** (Toulouse School of Economics)

Collaboration

Network analysis
(social, biological...)

Spatial statistics
(R package “GeoXp”)

Plan

Notations and examples

Data: A weighted undirected **network** modeled by a graph $\mathcal{G}$ with n nodes $x_1, \ldots, x_n$ with **weight matrix** W: $W_{ij} = W_{ji}$ and $W_{ii} = 0$.

Notations and examples

Data: A weighted undirected **network** modeled by a graph $\mathcal{G}$ with n nodes $x_1, \ldots, x_n$ with **weight matrix** W: $W_{ij} = W_{ji}$ and $W_{ii} = 0$.
For each node, **one or multiple label(s)** are given

$$\mathcal{C} : x_i \rightarrow \mathcal{C}(x_i) \subset \{c_1, \ldots, c_k\}$$

where c_j is either a numerical information or a factor information.

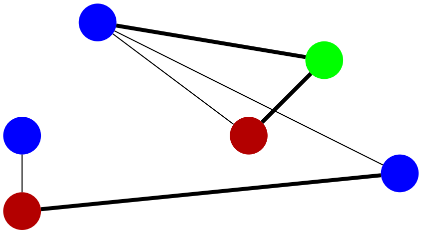
Notations and examples

Data: A weighted undirected **network** modeled by a graph $\mathcal{G}$ with n nodes $x_1, \ldots, x_n$ with **weight matrix** W: $W_{ij} = W_{ji}$ and $W_{ii} = 0$.

For each node, **one or multiple label(s)** are given

$$\mathcal{C} : x_i \rightarrow \mathcal{C}(x_i) \subset \{c_1, \ldots, c_k\}$$

where c_j is either a numerical information or a **factor information**.



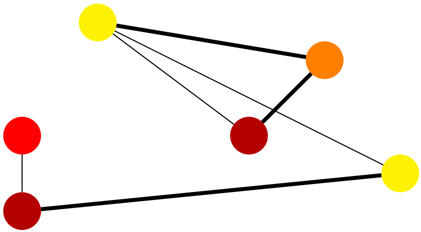
Examples: Gender in a social network, Functional group of a gene in a gene interaction network...

Notations and examples

Data: A weighted undirected **network** modeled by a graph $\mathcal{G}$ with n nodes $x_1, \ldots, x_n$ with **weight matrix** W: $W_{ij} = W_{ji}$ and $W_{ii} = 0$.
For each node, **one or multiple label(s)** are given

$$\mathcal{C} : x_i \rightarrow \mathcal{C}(x_i) \subset \{c_1, \ldots, c_k\}$$

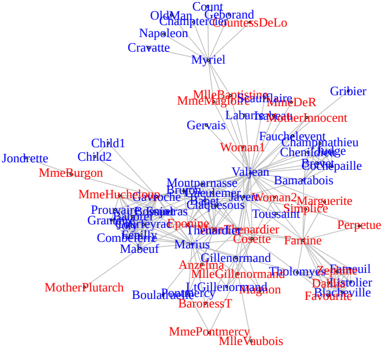
where c_j is either a **numerical information** or a factor information.

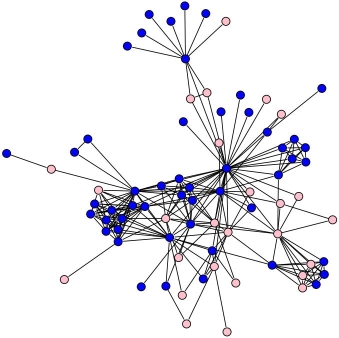


Examples: Weight of people in a social network, Number of visits of a web page in WWW...

"Real world" examples

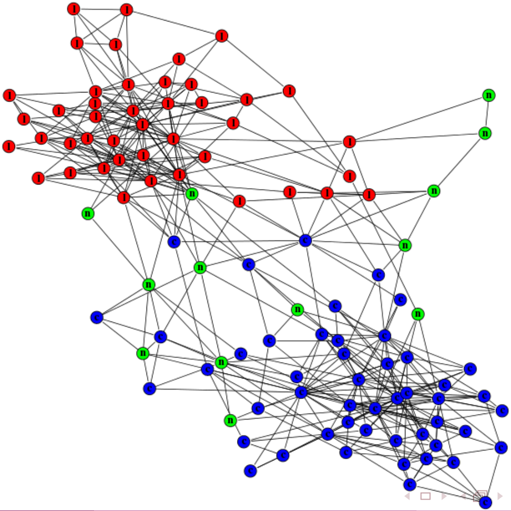
Example 1: Co-appearance network of the novel "Les Misérables" (Victor Hugo) where the nodes are labeled with gender (F/M).





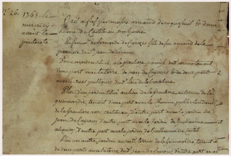
"Real world" examples

Example 2: Co-purchase network: nodes are books sold by "Amazon" and are labeled according to the political orientation of the book



"Real world" examples

Modeling a large corpus of medieval documents

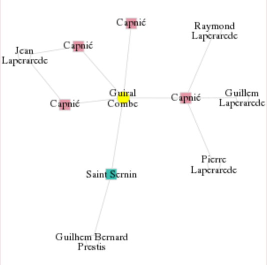


Notarial acts (mostly **baux à fief**, more precisely, land charters) established in a **seigneurie** named "Castelnau Montratier", written between 1250 and 1500, involving tenants and lords.[a]

[a] http://graphcomp.univ-tlse2.fr

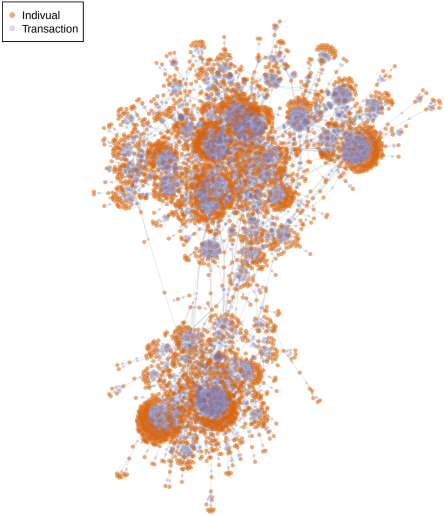
“Real world” examples

Modeling a large corpus of medieval documents



- nodes: transactions and individuals (3 918 nodes)
- edges: an individual is directly involved in a transaction (6 455 edges)
- labels (transactions only): location (parish)

"Real world" examples



Questions?

Is there a **link between the values of the nodes $(c_i)_i$ and the network structure**?

Questions?

Is there a **link between the values of the nodes $(c_i)_i$ and the network structure**?
Are the nodes labeled with a given value **more connected to nodes with the same value** than expected? less connected?
where *"expected"* means: in comparison to a random distribution over the network.

First approach: Use of "spatial" indexes

[Laurent and Villa-Vialaneix, 2011], by identifying

- the spatial matrix (in spatial data)
- the adjacency matrix (in network)

calculate

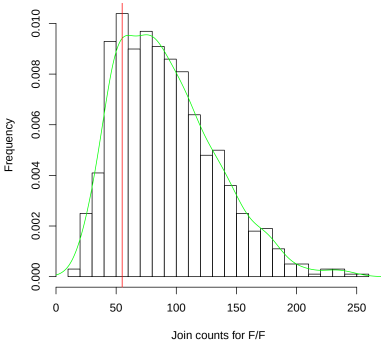
$$JC = \frac{1}{2} \sum_{i \neq j} W_{ij} \xi_i \xi_j$$

and a MC permutation test helps measuring the strength of the link between the labels and the network structure.

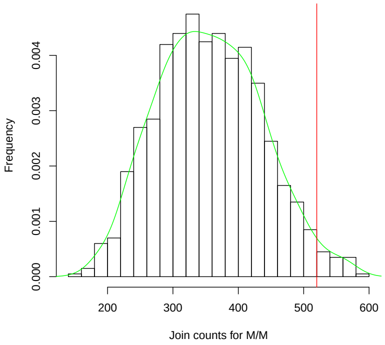
A toy example: “Les Misérables”

Data: Co-appearance network of the novel “Les Misérables” (Victor Hugo) where the nodes are labeled with gender (F/M).

Empirical distribution with Monte Carlo approach ($P = 1000$)



JC_F



JC_M

A toy example: “Les Misérables”

Data: Co-appearance network of the novel “Les Misérables” (Victor Hugo) where the nodes are labeled with gender (F/M).

Estimated p-value and conclusion

Gender	Join count value	Large	Small
F	55	0.7932 (NS)	0.2068 (NS)
M	520	0.0224 (**)	0.9755 (NS)

Men have a tendency to interact with other men rather than with women in “Les Misérables” whereas women don’t have a specific way to be related according to gender.

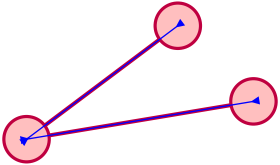
Plan

Alternative approach: graph visualization

Main idea: Find a representation of the graph that enlighten the labels information.

Alternative approach: graph visualization

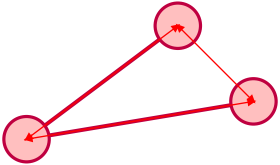
Main idea: Find a representation of the graph that enlighten the labels information. Graph visualization is a standard data mining tool to help the user understand the network. Standard approach are **force directed placement algorithms** as those introduced in **[Fruchterman and Reingold, 1991]**



- attractive forces : along the edges (similar to springs)

Alternative approach: graph visualization

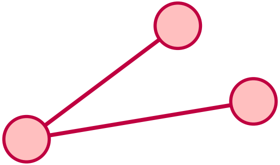
Main idea: Find a representation of the graph that enlighten the labels information. Graph visualization is a standard data mining tool to help the user understand the network. Standard approach are **force directed placement algorithms** as those introduced in **[Fruchterman and Reingold, 1991]**



- attractive forces : along the edges (similar to springs)
- repulsive forces : between all pairs of vertices (similar to electric forces)

Alternative approach: graph visualization

Main idea: Find a representation of the graph that enlighten the labels information. Graph visualization is a standard data mining tool to help the user understand the network. Standard approach are **force directed placement algorithms** as those introduced in **[Fruchterman and Reingold, 1991]**



- attractive forces : along the edges (similar to springs)
- repulsive forces : between all pairs of vertices (similar to electric forces)

iterative algorithm until stabilization of the nodes positions.

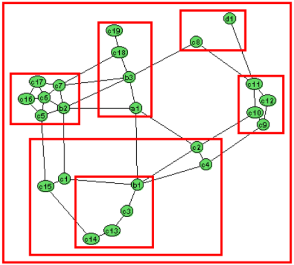
State-of-the art: clustered graph visualization

Main idea: Labels can be seen as a clustering ⇒ use visualization approach that allows **the nodes with the same labels to be displayed close to each others**.

State-of-the art: clustered graph visualization

Main idea: Labels can be seen as a clustering ⇒ use visualization approach that allows **the nodes with the same labels to be displayed close to each others**.

- **Modified force directed placement algorithms [Bourqui et al., 2007, Eades and Feng, 1996, Eades and Huang, 2000, Truong et al., 2007]**: integrate additional constraints into forces or constrain vertices to be displayed in a given zone, according to their clusters;



State-of-the art: clustered graph visualization

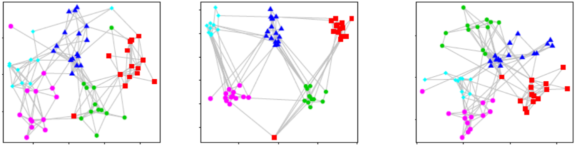
Main idea: Labels can be seen as a clustering ⇒ use visualization approach that allows **the nodes with the same labels to be displayed close to each others**.

- **Modified force directed placement algorithms**
  The graph can be displayed in a **simplified way** (one “meta-node” per cluster) as in **[Rossi and Villa-Vialaneix, 2011]**.

State-of-the art: clustered graph visualization

Main idea: Labels can be seen as a clustering ⇒ use visualization approach that allows **the nodes with the same labels to be displayed close to each others**.

- **Modified force directed placement algorithms**
- **Use of latent variables [Bouveyron et al., 2009]**



(a) Usual latent space (b) Supervised latent space (SL1) (c) Supervised latent space (SL2)

State-of-the art: clustered graph visualization

Main idea: Labels can be seen as a clustering ⇒ use visualization approach that allows **the nodes with the same labels to be displayed close to each others**.

- **Modified force directed placement algorithms**
- **Use of latent variables**

All these approaches:

- only consider the node's label and do not use the neighbors' labels;
- do not deal with multiple labels.

Plan

PCA based on neighbors' labels distribution

Denote:

- E the disjunctive encoding of nodes' labels

$$E_{ij} = \begin{cases} 1 & \text{if } c_j \in \mathcal{C}(x_i) \\ 0 & \text{if } c_j \notin \mathcal{C}(x_i) \end{cases}$$

PCA based on neighbors' labels distribution

Denote:

- E the disjunctive encoding of nodes' labels

$$E_{ij} = \begin{cases} 1 & \text{if } c_j \in \mathcal{C}(x_i) \\ 0 & \text{if } c_j \notin \mathcal{C}(x_i) \end{cases}$$

- P_l, the labels distribution among the neighbors:

$$P_l = D^{-1} W E$$

where $D = \mathrm{Diag}(d_1, \ldots, d_n)$ with d_i degree of node x_i.

PCA based on neighbors' labels distribution

Denote:

- E the disjunctive encoding of nodes' labels

$$E_{ij} = \begin{cases} 1 & \text{if } c_j \in \mathcal{C}(x_i) \\ 0 & \text{if } c_j \notin \mathcal{C}(x_i) \end{cases}$$

- P_l, the labels distribution among the neighbors:

$$P_l = D^{-1} W E$$

where $D = \mathrm{Diag}(d_1, \ldots, d_n)$ with d_i degree of node x_i.

Display the graph with the coordinates in the **Weighted PCA** of P_l where columns are weighted by $\frac{n_j}{n}$ with $n_j = \left|\{x_i :\ c_j \in \mathcal{C}(x_i)\}\right|$.

PCA based on neighbors' labels distribution

Denote:

- E the disjunctive encoding of nodes' labels

$$E_{ij} = \begin{cases} 1 & \text{if } c_j \in \mathcal{C}(x_i) \\ 0 & \text{if } c_j \notin \mathcal{C}(x_i) \end{cases}$$

- P_l, the labels distribution among the neighbors:

$$P_l = D^{-1} W E$$

where $D = \mathrm{Diag}(d_1, \ldots, d_n)$ with d_i degree of node x_i.

Display the graph with the coordinates in the **Weighted PCA** of P_l where columns are weighted by $\frac{n_j}{n}$ with $n_j = \left|\{x_i : \ c_j \in \mathcal{C}(x_i)\}\right|$.

Remark: This choice is similar to the use of the χ^2 metric:

$$\delta(p_i, p_{i'}) = \sum_c \frac{n}{n_c} \left(\frac{n_{ic}}{d_i} - \frac{n_{i'c}}{d_{i'}} \right)^2$$

Kernel based approach

Previous method **drawbacks**:

- do not use the label of the node but only those of its neighbors;
- only use the direct neighbors' labels.

Kernel based approach

Previous method **drawbacks**:

- do not use the label of the node but only those of its neighbors;
- only use the direct neighbors' labels.

Alternative approach: Use a diffusion process by means of the **heat kernel**

$$K^{\beta} = e^{-\beta L}$$

where $L = D - W$.

Kernel based approach

Previous method **drawbacks**:

- do not use the label of the node but only those of its neighbors;
- only use the direct neighbors' labels.

Alternative approach: Use a diffusion process by means of the **heat kernel**

$$K^\beta = e^{-\beta L}$$

where $L = D - W$.

Heat kernel features:

- has a simple interpretation regarding a diffusion process along the edges of the graph;
- can be viewed as a dot product between nodes in an embedding space:

$$K_{ij}^\beta \equiv K^\beta(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{K}^\beta}$$

Kernel PCA for labeled graph visualization

- Use $K^\beta E$ instead of P_l to represent the labels distribution among the neighbors (the node's label is used):

$$\tilde{f}_{ic}^{\beta} = \langle \phi(x_i), \sum_{j:c_j=c} \phi(x_j) \rangle_{\mathcal{K}^\beta}$$

Kernel PCA for labeled graph visualization

- Use $K^\beta E$ instead of P_l to represent the labels distribution among the neighbors (the node's label is used):

$$\tilde{f}^\beta_{ic} = \langle \phi(x_i), \sum_{j:c_j=c} \phi(x_j) \rangle_{\mathcal{K}^\beta}$$

- Display the graph with the coordinates in the **Weighted PCA** of $K^\beta E$ where columns are weighted by $\frac{n_j}{n}$ with $n_j = \left|\{x_i :\ c_j \in \mathcal{C}(x_i)\}\right|$.

Kernel PCA for labeled graph visualization

- Use $K^\beta E$ instead of P_l to represent the labels distribution among the neighbors (the node's label is used):

$$\tilde{f}_{ic}^{\beta} = \langle \phi(x_i), \sum_{j:c_j=c} \phi(x_j) \rangle_{\mathcal{K}^\beta}$$

- Display the graph with the coordinates in the **Weighted PCA** of $K^\beta E$ where columns are weighted by $\frac{n_j}{n}$ with $n_j = \left|\{x_i : c_j \in \mathcal{C}(x_i)\}\right|$.

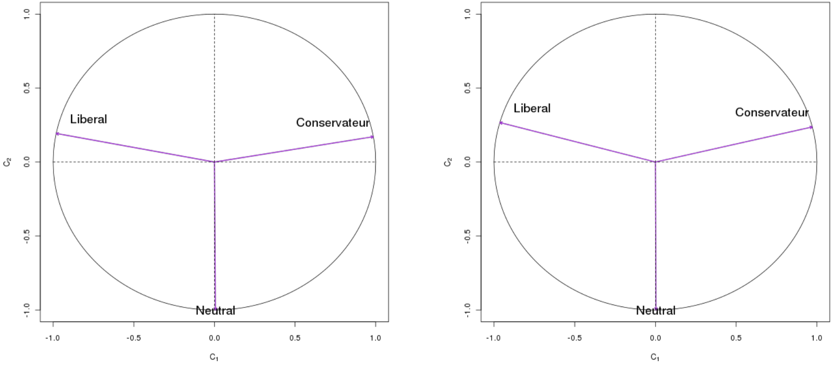
Various β will provide various representation: small β favor direct neighbors.

Plan

Polbooks

Co-purchase network: nodes are books sold by “Amazon” and are labeled according to the political orientation of the book.

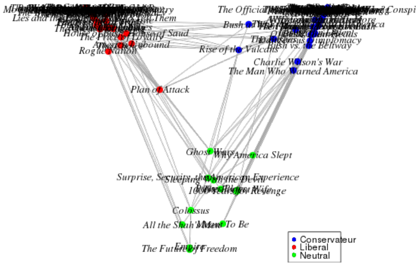
Labels representation



Polbooks

Co-purchase network: nodes are books sold by “Amazon” and are labeled according to the political orientation of the book.

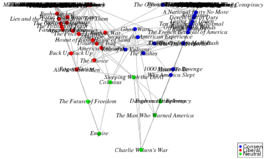
Network representation

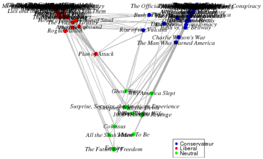


Polbooks

Co-purchase network: nodes are books sold by “Amazon” and are labeled according to the political orientation of the book.

Network representation





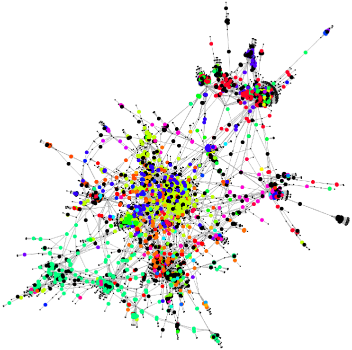
Main conclusions:

- Strong relations between labels and graph structure: nodes with the same labels also have the same labels distribution among their respective neighbors;
- Provide a **more subtle interpretation** of the book’s political orientation (ex: “World of Vulcain” is conservative but close to liberal)
- Differences between the two representations (ex: “Plan of attack” is frequently co-purchased with liberal books that are themselves frequently co-purchased with non-liberal books)

Medieval

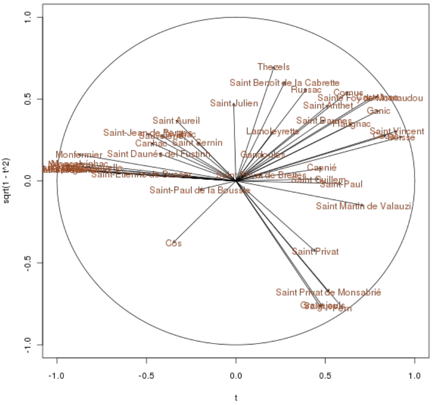
Bipartite graph

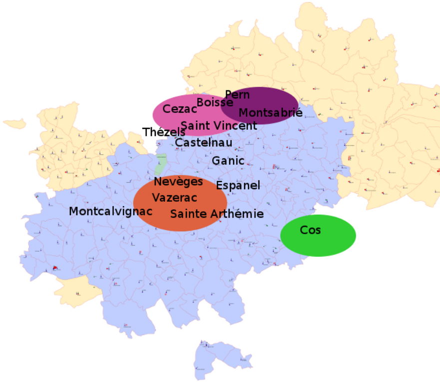
- nodes: transactions and individuals (3 918 nodes)
- edges: an individual is directly involved in a transaction (6 455 edges)
- labels (transactions only): location (parish)



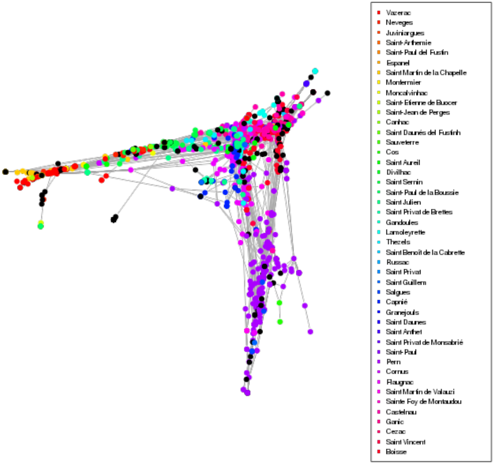
Medieval

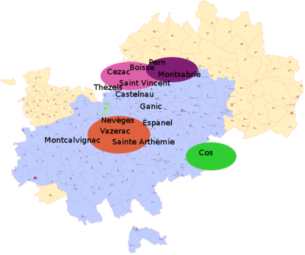
PCA applied on the **individuals** only (projected network) based on the location distribution among transactions (multiple labels).





Medieval


Vazerac
Neveges
Juviniargues
Saint-Arthemie
Saint-Paul del Fustin
Espanel
Saint Martin de la Chapelle
Montermier
Moncalvinhac
Saint-Etienne de Bucer
Saint-Jean de Perges
Canhac
Saint Daunès del Fustinh
Sauveterre
Cos
Saint Aureil
Divilhac
Saint Sernin
Saint-Paul de la Boussie
Saint Julien
Saint Privat de Brettes
Gandoules
Lamoleyrette
Thezels
Saint Benoît de la Cabrette
Russac
Saint Privat
Saint Guillem
Salgues
Caprié
Granejouls
Saint Daunes
Saint Anthet
Saint Privat de Monsabrié
Saint-Paul
Pern
Cornus
Flaugnac
Saint Martin de Valauzi
Sainte Foy de Montaudou
Castelnau
Ganic
Cezac
Saint Vincent
Boisse


Pern
Cezac
Boisse
Montsabrié
Thézels
Saint Vincent
Castelnau
Ganic
Nevèges
Espanel
Vazerac
Montcalvignac
Sainte Arthémie
Cos

Any questions?...

Bourqui, R., Auber, D., and Mary, P. (2007).
How to draw clustered weighted graphs using a multilevel force-directed graph drawing algorithm.
In *Proceedings of the 11th International Conference Information Visualization, 2007. IV'07.*, pages 757–764.

Bouveyron, C., Chipman, H., and Côme, E. (2009).
Supervised classification and visualization of social networks based on a probabilistic latent space model.
In *Proceedings of 7th International Workshop on Mining and Learning with Graphs*, Leuven, Belgium.

Eades, P. and Feng, Q. (1996).
Multilevel visualization of clustered graphs.
In North, S. C., editor, *Proceedings of International Conference on Graph Drawing, Symposium on Graph Drawing*, volume 1190 of *Lecture Notes in Computer Science*, pages 101–112, Berkeley, California, USA. Springer.

Eades, P. and Huang, M. (2000).
Navigating clustered graphs using force-directed methods.
Journal of Graph Algorithms and Applications, 4(3):157–181.

Fruchterman, T. and Reingold, B. (1991).
Graph drawing by force-directed placement.
Software-Practice and Experience, 21:1129–1164.

Laurent, T. and Villa-Vialaneix, N. (2011).
Using spatial indexes for labeled network analysis.
Information, Interaction, Intelligence (i3), 11(1).

Rossi, F. and Villa-Vialaneix, N. (2011).
Représentation d'un grand réseau à partir d'une classification hiérarchique de ses sommets.
Journal de la Société Française de Statistique, 152(3):34–65.

Truong, Q., Dkaki, T., and Charrel, P. (2007).
An energy model for the drawing of clustered graphs.
In *Proceedings of Vème colloque international VSST*, Marrakech, Maroc.