Permutation tests for labeled network analysis

Nathalie Villa-Vialaneix

http://www.nathalievilla.org

nathalie.villa@univ-paris1.fr

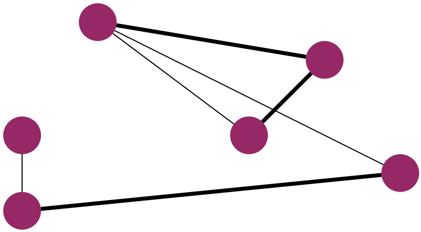
ERCIM 2013, London, 14-16 December 2013

Joint work with *Thibault Laurent* (Toulouse School of Economics) *Bertrand Jouve* (Université Lyon 2), *Fabrice Rossi* (Université Paris 1) & *Florent Hautefeuille* (Université Toulouse 2)

Outline

Framework

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



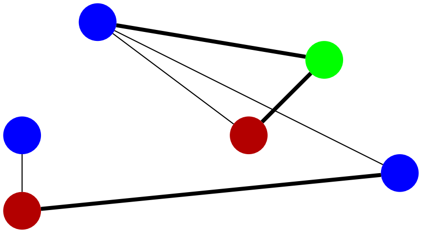
Used to represent **relations between entities** (social network, gene regulation network, ...)

Framework

Data: A weighted undirected **network/graph** $\mathcal{G}$ with n **nodes** $x_1, \ldots, x_n$ and **weight matrix** W st: $W_{ij} = W_{ji} \geq 0$ and $W_{ii} = 0$.
For each node, an **additional information**

$$\mathcal{C} : x_i \rightarrow c_i$$

c_i: numeric ($c_i \in \mathbb{R}$), factor ($c_i \in \{m_1, \ldots, m_k\}$) or spatial information.



Examples: Genders in a social network, Functional groups genes in a gene regulation network, Weight of people in a social network, Number of visits of a web page in WWW, Place of home in a social network...

Questions?

Is there a link between the node labels $(c_i)_i$ and the network structure?

Questions?

Is there a link between the node labels $(c_i)_i$ and the network structure?

- For a **factor label**, are the nodes labelled with a given value more connected to nodes with the same value than expected? less connected?
- For a **numerical label**, are the numerical values of the nodes more correlated to the values of connected nodes than expected?
- For a **spatial label**, is there a stronger/smaller proximity than expected between the spatial labels of connected nodes?

where “expected” means: in comparison to a random distribution over the network.

Analogy between spatial statistics and network analysis

Spatial statistics: spatial units $(x_i)_i$ frequently described by a spatial matrix W st W_{ij} encodes adjacency between x_i and x_j (sometimes row/column normalized)

Network analysis: nodes $(x_i)_i$ described by a neighbourhood matrix W, which is symmetric

Outline

Join Count Statistics

Binary labels: $c_i \in \{0, 1\}$.
General form:

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

where ξ_i is either c_i or $1 - c_i$.

Join Count Statistics

Binary labels: $c_i \in \{0, 1\}$.
General form:

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

where ξ_i is either c_i or $1 - c_i$.
Basic interpretation: JC_1 "large" (/"small") $\Leftrightarrow$ nodes labelled "1" tends to be linked to nodes labelled the same way (/the opposite way)

Join Count Statistics

Binary labels: $c_i \in \{0, 1\}$.
General form:

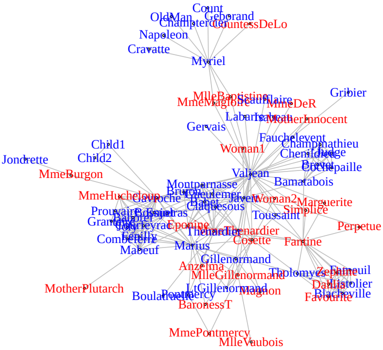
$$JC = \frac{1}{2} \sum_{i \neq j} W_{ij} \xi_i \xi_j$$

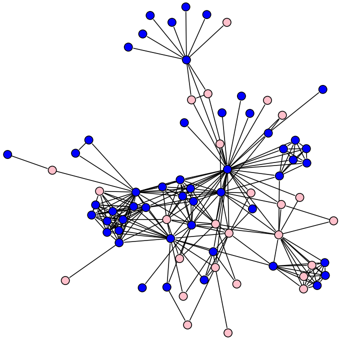
where ξ_i is either c_i or $1 - c_i$.
Basic interpretation: JC_1 "large" (/"small") $\Leftrightarrow$ nodes labelled "1" tends to be linked to nodes labelled the same way (/the opposite way)
Statistical significance: When is JC_1 significantly large or small?

- **Method 1**: [Noether, 1970] JC_1 is asymptotically normally distributed but requires additional assumptions on the network structure and not valid for small networks;
- **Method 2: Monte Carlo approach**: Randomly permute c_i over the nodes, P times $\Rightarrow$ empirical distribution of JC_1 compared to the actual JC_1.

A toy example: “Les Misérables”

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

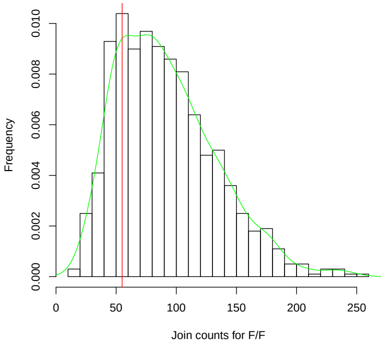




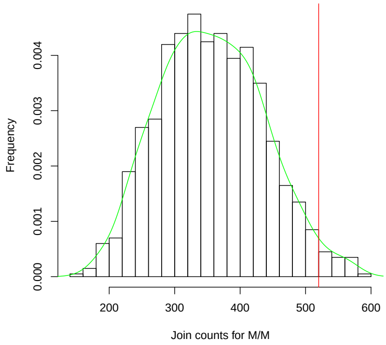
A toy example: “Les Misérables”

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

Estimated p-value and conclusion

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

Moran's I

Numeric labels: $c_i \in \mathbb{R}$.
[Moran, 1950], ***I* statistics**:

$$I = \frac{\frac{1}{2m}\sum_{i \neq j} W_{ij}\bar{c}_i\bar{c}_j}{\frac{1}{n}\sum_i \bar{c}_i^2}$$

where $m = \frac{1}{2}\sum_{i \neq j} W_{ij}$ and $\bar{c}_i = c_i - \bar{c}$ with $\bar{c} = \frac{1}{n}\sum_i c_i$.

Moran's I

Numeric labels: $c_i \in \mathbb{R}$.
[Moran, 1950], ***I* statistics**:

$$I = \frac{\frac{1}{2m}\sum_{i\neq j} W_{ij}\bar{c}_i\bar{c}_j}{\frac{1}{n}\sum_i \bar{c}_i^2}$$

where $m = \frac{1}{2}\sum_{i\neq j} W_{ij}$ and $\bar{c}_i = c_i - \bar{c}$ with $\bar{c} = \frac{1}{n}\sum_i c_i$.
Interpretation: I "large" $\Leftrightarrow$ nodes tend to be connected to nodes which have similar labels

Moran's I

Numeric labels: $c_i \in \mathbb{R}$.
[Moran, 1950], ***I* statistics**:

$$I = \frac{\frac{1}{2m}\sum_{i\neq j} W_{ij}\bar{c}_i\bar{c}_j}{\frac{1}{n}\sum_i \bar{c}_i^2}$$

where $m = \frac{1}{2}\sum_{i\neq j} W_{ij}$ and $\bar{c}_i = c_i - \bar{c}$ with $\bar{c} = \frac{1}{n}\sum_i c_i$.
Interpretation: I "large" $\Leftrightarrow$ nodes tend to be connected to nodes which have similar labels
Deriving a test for *I*: once again, **asymptotic normality can be proved** but using a **Monte Carlo simulation** is useful for small network cases.

Outline

Relational data coming from a large corpus of medieval documents 1/2

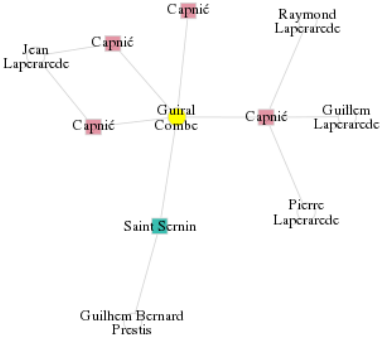
A large corpus of notarial acts

The corpus has been re-written by a feudist during the XIXe century and is kept at the **archives départementales du Lot (Cahors, France)**

- notarial acts related to rents (mostly “baux à fief”);
- established between 1250 and 1500;
- in the seigneurie (about 10 little villages) called Castelnau Montratier (Lot, France).

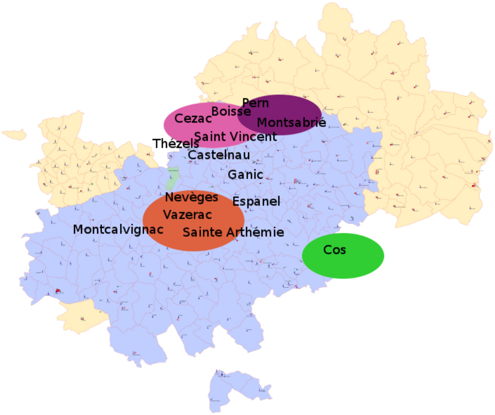
Relational data coming from a large corpus of medieval documents 2/2

Defining a **bipartite graph**:



- nodes: transactions and individuals (3 918 nodes)
- edges: an individual directly involved in a transaction (6 455 edges)
- labels: for individuals (name, role...), for transactions (place: parish, date...)

Spatial labels



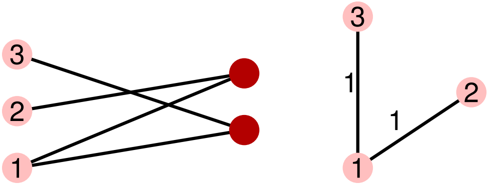
transactions are spatially localized: 45 parishes (known positions);

Question: What is the impact of the spatial locations of lands exchanged in the transactions on the way the individuals interact?

Graphs built from medieval documents

From the bipartite graph, define a **projected graph**:

- nodes are the individuals
- an edge connects two individuals if they are involved in the same transaction (edges can eventually be weighted)



Social distances between individuals $(S_{ij})_{ij}$

Quantifying the interactions between individuals

Idea: Use the previous graph as a measure of social distance.

Social distances between individuals $(S_{ij})_{ij}$

Quantifying the interactions between individuals

Idea: Use the previous graph as a measure of social distance.

Used dissimilarities/similarities:

1. **shortest path length** on the graph;

Social distances between individuals $(S_{ij})_{ij}$

Quantifying the interactions between individuals

Idea: Use the previous graph as a measure of social distance.

Used dissimilarities/similarities:

1. **shortest path length** on the graph;
2. **similarities based on the adjacency matrix**:

$$A_{ij} = \begin{cases} 1 & \text{si les sommets } i \text{ et } j \text{ sont liés par une arête} \\ 0 & \text{sinon.} \end{cases}$$

Social distances between individuals $(S_{ij})_{ij}$

Quantifying the interactions between individuals

Idea: Use the previous graph as a measure of social distance.

Used dissimilarities/similarities:

1. **shortest path length** on the graph;
2. **similarities based on the adjacency matrix**: **regularized versions of the Laplacian** ($L = \mathrm{Diag}(d_i)_i - A$ where d_i is the degree of node i):

$$L^+$$

e.g., “commute time kernel” **[Fouss et al., 2007]**

Spatial distances between individuals $(G_{ij})_{ij}$

Quantifying spatial distances between individuals $(G_{ij})_{ij}$

Idea:

- List of parishes cited in the transactions in which the individual is involved;
- Center of gravity of these locations.

Spatial distances between individuals $(G_{ij})_{ij}$

Quantifying spatial distances between individuals $(G_{ij})_{ij}$

Idea:

- List of parishes cited in the transactions in which the individual is involved;
- Center of gravity of these locations.

Spatial distance: distance between centers of gravity.

Standard approach: Mantel's test

Test of the correlation between two matrices S and G of distances

As distances are not independent data, use **Mantel's test**:

- permute P times the rows and columns of S;
- compute the corresponding correlation coefficients: $\mathrm{Cor}^p(S^p, G^p)$.

and use it as an empirical distribution for the independence between S and G.

Standard approach: Mantel's test

Test of the correlation between two matrices S and G of distances

As distances are not independent data, use **Mantel's test**:

- permute P times the rows and columns of S;
- compute the corresponding correlation coefficients: $\mathrm{Cor}^p(S^p, G^p)$.

and use it as an empirical distribution for the independence between S and G.

But (here) S and G are built from the same network: they are correlated and permuting rows and columns of G does not respect the dependency structure and is not the empirical distribution of the null hypothesis:

social distances are not related to spatial locations

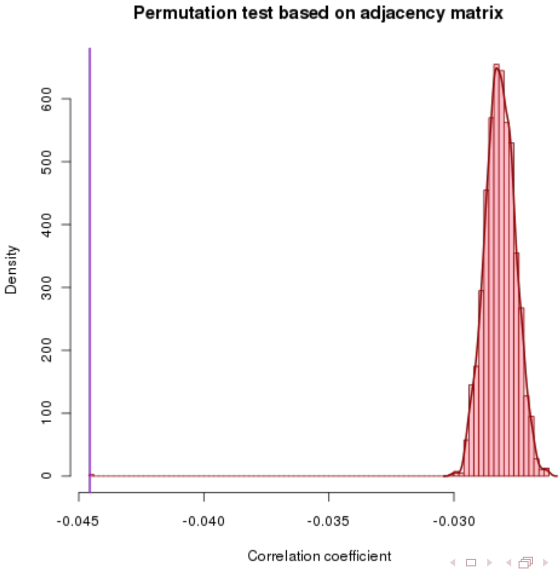
Adapting Mantel's test

Permutation test based on the bipartite graph

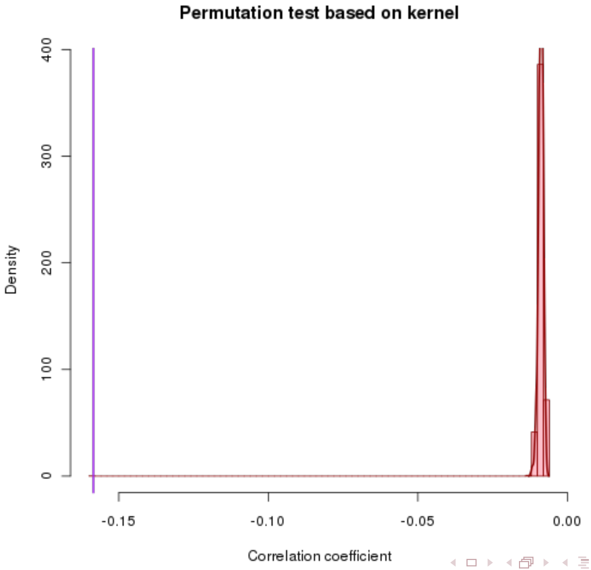
Repeat P times

1. permute spatial labels between transactions (empirical distribution of the null hypothesis on spatial labels);
2. compute the corresponding $(G^p_{ij})_{ij}$;
3. compute the corresponding correlation coefficient Cor^p between S and G^p.

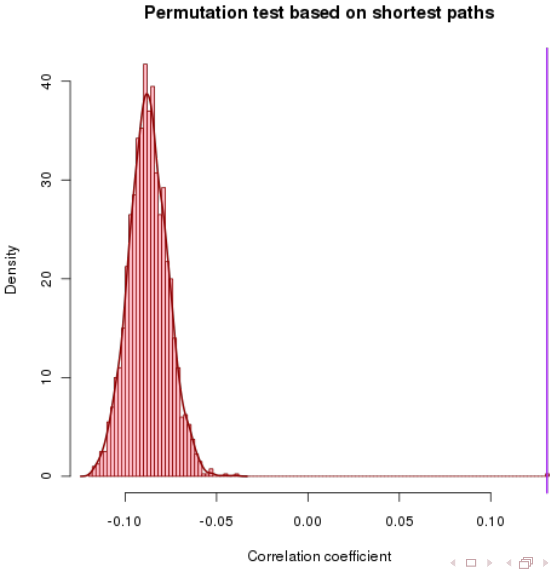
Results obtained with various social distances



Results obtained with various social distances



Results obtained with various social distances



Conclusion

- Spatial indexes can help describe and analyze the distribution of a given variable on the nodes of a network;
- Permutation tests can be used to analyze the correlation between the network structure and node labels for various types of labels.

Related work:
[Laurent and Villa-Vialaneix, 2011, Villa-Vialaneix et al., 2012]

Thank you for your attention... Any question?

Fouss, F., Pirotte, A., Renders, J., and Saerens, M. (2007).
Random-walk computation of similarities between nodes of a graph, with application to collaborative recommendation.
IEEE Transactions on Knowledge and Data Engineering, 19(3):355–369.

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

Moran, P. (1950).
Notes on continuous stochastic phenomena.
Biometrika, 37:17–23.

Noether, G. (1970).
A central limit theorem with non-parametric applications.
Annals of Mathematical Statistics, 41:1753–1755.

Villa-Vialaneix, N., Jouve, B., Rossi, F., and Hautefeuille, F. (2012).
Spatial correlation in bipartite networks: the impact of the geographical distances on the relations in a corpus of medieval transactions.
Revue des Nouvelles Technologies de l'Information, SHS-1:97–110.