Analysis of the influence of a network on the values of its nodes: the use of spatial indexes

Nathalie Villa-Vialaneix[♀]
http://www.nathalievilla.org

& Thibaut Laurent (TSE)

[♀] IUT de Carcassonne (UPVD)

& Institut de Mathématiques de Toulouse

MARAMI 2010

Toulouse, 11/12 octobre 2010

Collaboration

Network analysis
(social, biological...)

Spatial statistics
(R package “GeoXp”)

Notations and examples

Data: A weighted undirected **network** modelled 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** modelled 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, an **additional information**

$$C : x_i \to c_i$$

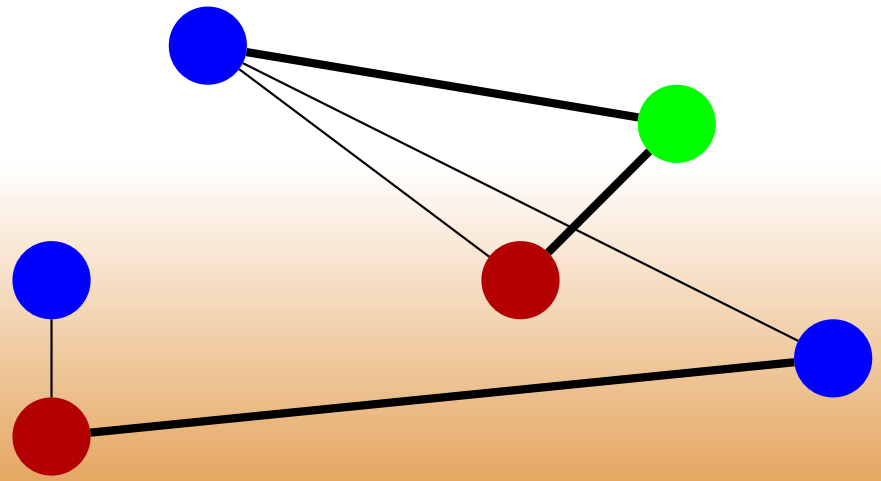
where c_i is either a numerical information or a factor information.

Notations and examples

Data: A weighted undirected **network** modelled 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, an **additional information**

$$C : x_i \to c_i$$

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



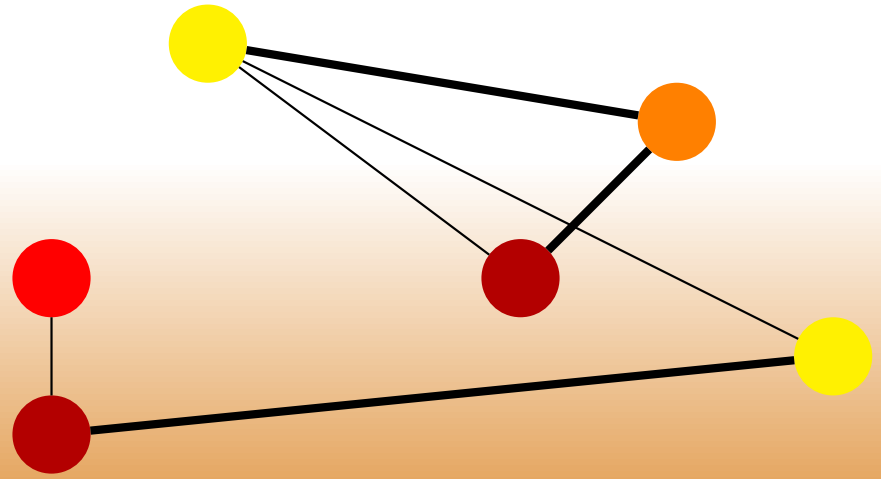
Examples: Gender in a social network, Functional group of a gene in a gene interaction network...

Notations and examples

Data: A weighted undirected **network** modelled 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, an **additional information**

$$C : x_i \to c_i$$

where c_i 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...

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?

- For a **factor information**, are the nodes labelled 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.

Questions?

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

- For a **factor information**, are the nodes labelled with a given value more connected to nodes with the same value than expected? less connected?
- For a **numerical information**, are the numerical values of the nodes more correlated to the values of the connected nodes than expected?

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

Questions?

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

- For a **factor information**, are the nodes labelled with a given value more connected to nodes with the same value than expected? less connected?
- For a **numerical information**, are the numerical values of the nodes more correlated to the values of the connected nodes than expected?

where “expected” means: in comparison to a random distribution over the network.
Use of **spatial indexes** by identifying

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

Warning!

This approach is not a **diffusion model**:

- in **diffusion models**, one tries to understand the way the information **propagates through time** over a network (modelization approach);
- **here**, we try to understand if a given (static) network is related to the values of its nodes (**descriptive approach**).

Outline

Join Count Statistics

Binary information: $c_i \in \{0, 1\}$.

Join Count Statistics

Binary information: $c_i \in \{0, 1\}$.

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 information: $c_i \in \{0, 1\}$.
Derived statistics:

- Number of "1" labels in the neighbor of a node labelled "1"

$$JC_1 = \frac{1}{2} \sum_{i,j:\ c_i=c_j=1} W_{ij}$$

- Number of "0" labels in the neighbor of a node labelled "0"

$$JC_0 = \frac{1}{2} \sum_{i,j:\ c_i=c_j=0} W_{ij}$$

- Number of "1" labels in the neighbor of a node labelled "0" (and the opposite)

$$JC_{0-1} = \sum_{i,j:\ c_i=0,\ c_j=1} W_{ij}$$

Interpretation

Basic interpretation: If JC_1 is “large” (“small”) then nodes labelled “1” in the network tends to be related with nodes labelled the same way (or tends not to be related to nodes labelled the same way).

Interpretation

Basic interpretation: If JC_1 is "large" ("small") then nodes labelled "1" in the network tends to be related with nodes labelled the same way (or tends not to be related to nodes labelled the same way).
Statistical significance: When is JC_1 significantly large or small?

- **Method 1**: **[Noether, 1970]** proves the asymptotic normal distribution of JC_1: requires additionnal assumptions on the network and not valid for small networks;

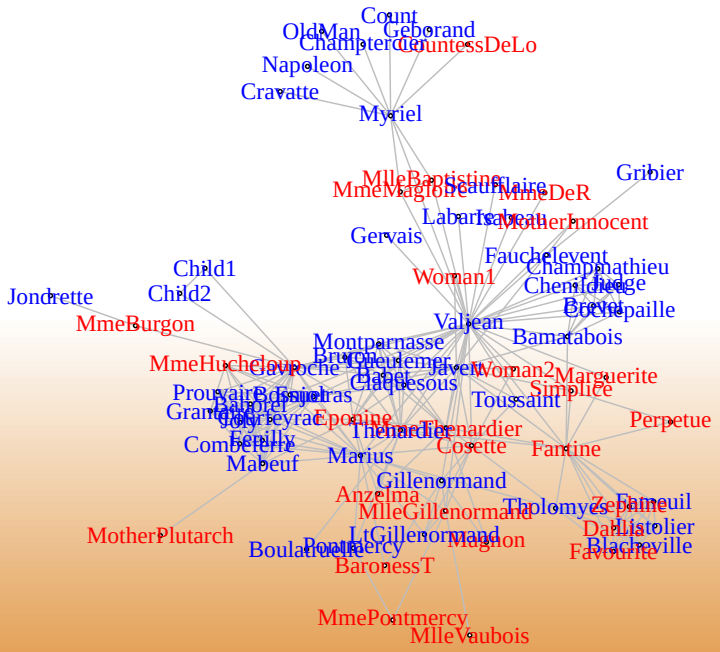
Interpretation

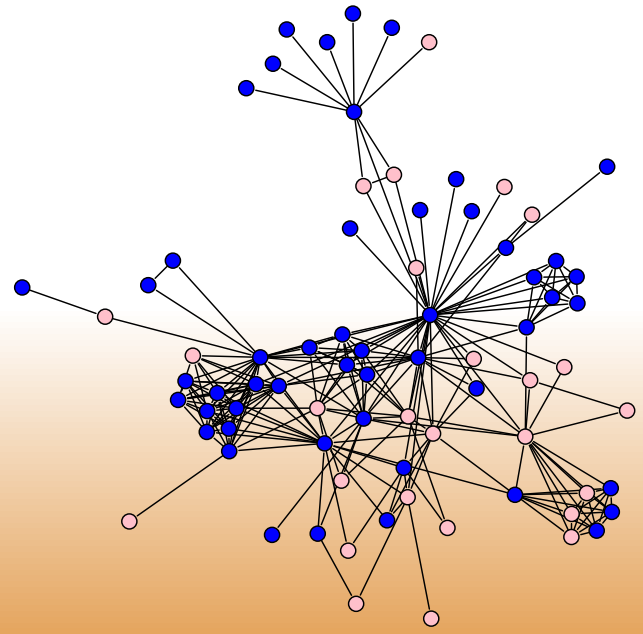
Basic interpretation: If JC_1 is "large" ("small") then nodes labelled "1" in the network tends to be related with nodes labelled the same way (or tends not to be related to nodes labelled the same way).
Statistical significance: When is JC_1 significantly large or small?

- **Method 1**: **[Noether, 1970]** proves the asymptotic normal distribution of JC_1: requires additionnal assumptions on the network and not valid for small networks;
- **Method 2: Monte Carlo approach**: Randomly permutate the values c_i over the nodes, P times (where P is large) and obtain the empirical distribution of JC_1. Compare with the true JC_1.
  $\Rightarrow$ Estimation of the distribution of JC_1 given the network and the numbers of "1" and "0" labels.

A toy example: “Les Misérables”

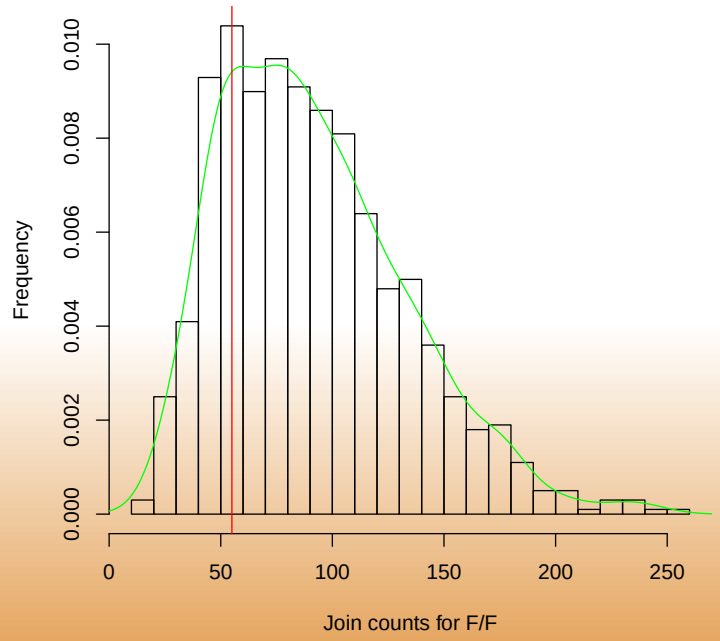
Data: Co-appearance network of the novel “Les Misérables” (Victor Hugo) where the nodes are labelled with gender (F/M).



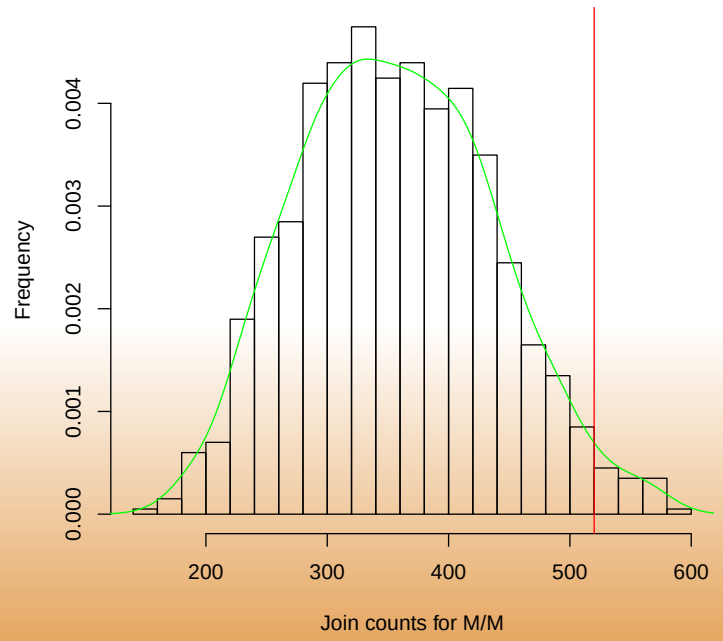


A toy example: “Les Misérables”

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 (NS)	0.2068 (NS)
M	520	0.0224 (**)	0.9755 (NS)

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 (NS)	0.2068 (NS)
M	520	0.0224 (**)	0.9755 (NS)

Men have a tendancy 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.

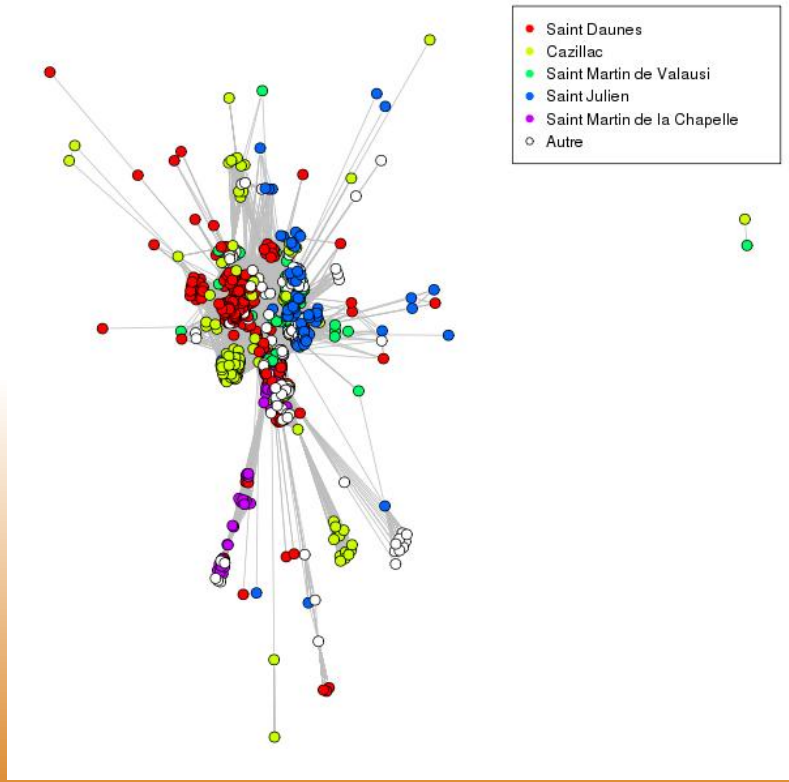
Example 2: Location information in a medieval social network

Basic description of the origin of the network: Built from agrarian contracts established between 1295 and 1336 (before the Hundred Years' War)[1]:

- **vertices**: peasants;
- **edges**: number of contracts shared by two peasants;
- **labels of the nodes**: main geographical location recorded for the peasant (5 main locations + "other")

Example 2: Location information in a medieval social network

Basic description of the origin of the network: Built from agrarian contracts established between 1295 and 1336 (before the Hundred Years' War)[1]:



[1] more details in **[Bouiet et al., 2008]**

Join count test conclusions

"One-against-all" approach: For each village, test $C :=$ belonging or not to this village.

Join count test conclusions

"One-against-all" approach: For each village, test $C :=$ belonging or not to this village.

Weighted network:

Location	Join count value	Large	Small
Saint-Daunes	110 892	0.0010 (***)	0.999 (NS)
Cazillac	24 461	0.0010 (***)	0.999 (NS)
Saint-Martin de Valausi	19 996	0.0010 (***)	0.999 (NS)
Saint-Julien	1 172	0.988 (NS)	0.0120 (**)
Saint-Martin de la Chapelle	10 200	0.0010 (***)	0.999 (NS)

Join count test conclusions

"One-against-all" approach: For each village, test $C :=$ belonging or not to this village.

Weighted network:

Location	Join count value	Large	Small
Saint-Daunes	110 892	0.0010 (***)	0.999 (NS)
Cazillac	24 461	0.0010 (***)	0.999 (NS)
Saint-Martin de Valausi	19 996	0.0010 (***)	0.999 (NS)
Saint-Julien	1 172	0.988 (NS)	0.0120 (**)
Saint-Martin de la Chapelle	10 200	0.0010 (***)	0.999 (NS)

Part of the explanation: St-Julien corresponds to three different villages that have the same name.

Outline

Moran's I

[Moran, 1950] proposes to measure spatial correlation with the ***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

[Moran, 1950] proposes to measure spatial correlation with the ***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: When I is "large", nodes tend to be connected to other nodes having close values for C; when I is "small", nodes tend to be connected to other nodes having very different values for C. Average I means that there is no special relation between C and the relations in the network.

Moran's I

[Moran, 1950] proposes to measure spatial correlation with the ***I* statistics**:

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

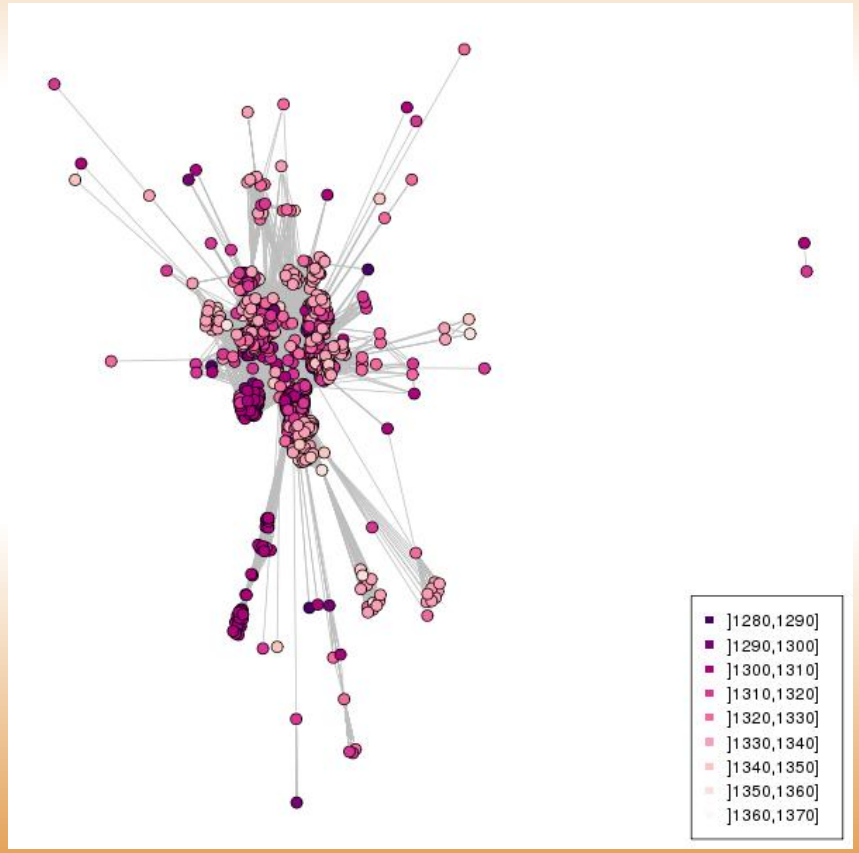
where $m = \frac{1}{2}\sum_{i\neq j} W_{ij}$ and $\bar{c}_i = c_i - \bar{c}$ with $\bar{c} = \frac{1}{n}\sum_i c_i$.

Interpretation: When I is "large", nodes tend to be connected to other nodes having close values for C; when I is "small", nodes tend to be connected to other nodes having very different values for C. Average I means that there is no special relation between C and the relations in the network.

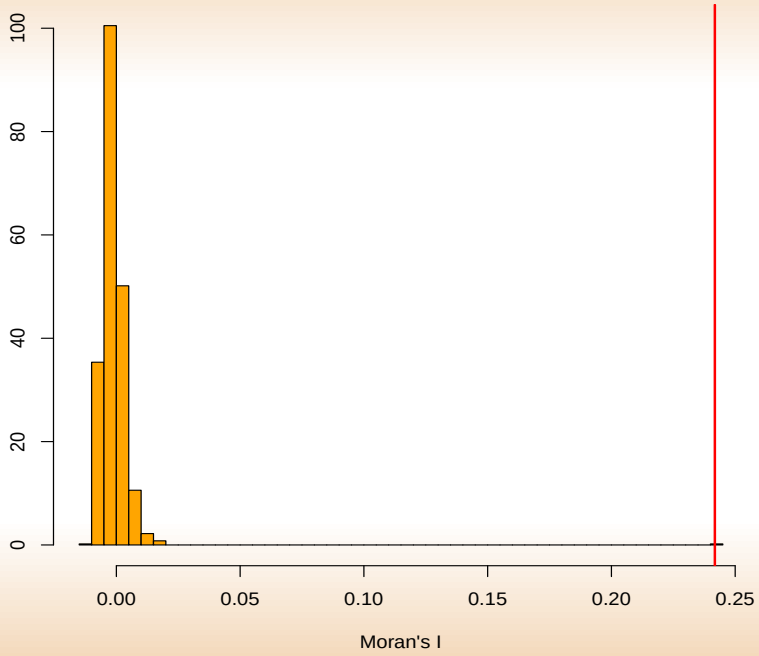
Deriving a test for *I*: once again, **asymptotic normality can be proved** but we prefer using a **Monte Carlo simulation** to estimate the distribution of I on our network.

Example 1: Dates in the medieval social network

Medians of the dates of activity for each peasant (1280-1360)



Example 1: Dates in the medieval social network



p-value for testing the hypothesis "I is large": $< 0.1\% \Rightarrow$ **Peasants tend to be connected to people having a very similar date of activity.**

Example 2: Relation between a gene co-expression network and an interesting phenotype

Data: Genes co-expression network from pig muscles:

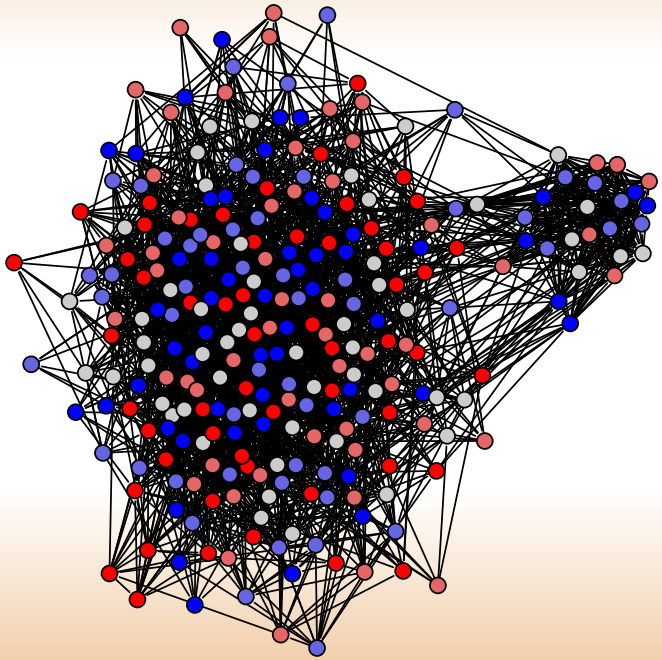
- **nodes** (272): genes those expression has been measured and that have been selected from the original sets of 2464 genes because they are **regulated by an eQTL** (genetic selection);
- **edges**: does a pair of gene have a highly correlated expression over the 57 pigs? (weighted by the **partial correlation**);

Example 2: Relation between a gene co-expression network and an interesting phenotype

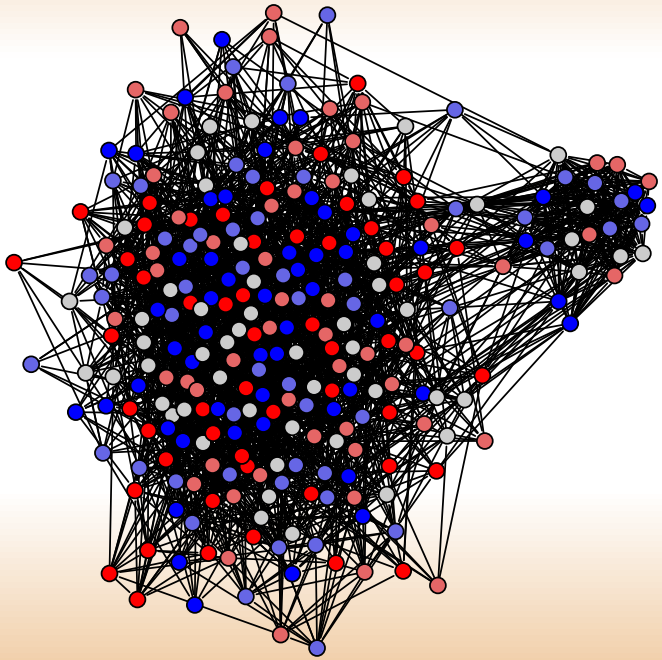
Data: Genes co-expression network from pig muscles:

- **nodes** (272): genes those expression has been measured and that have been selected from the original sets of 2464 genes because they are **regulated by an eQTL** (genetic selection);
- **edges**: does a pair of gene have a highly correlated expression over the 57 pigs? (weighted by the **partial correlation**);
- **additional information about the genes**: partial correlation with an interesting phenotype involved in the meat quality (PH).

Example 2: Relation between a gene co-expression network and an interesting phenotype



Example 2: Relation between a gene co-expression network and an interesting phenotype

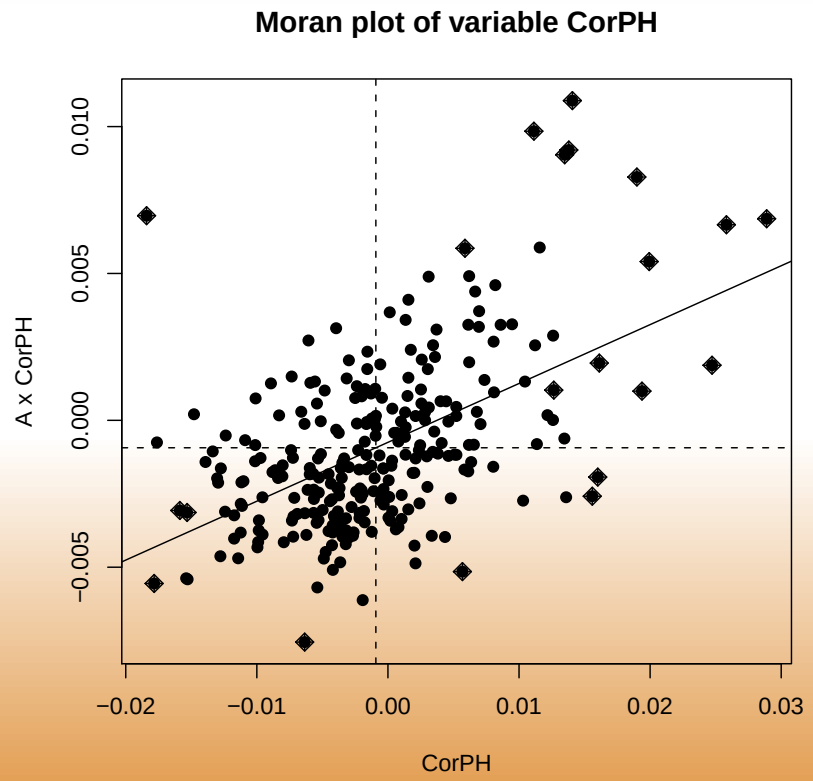


p-value for testing the hypothesis "I is large": < 0.1

$\Rightarrow$ **related genes tends to share the same kind of correlation with PH**.

Further analysis

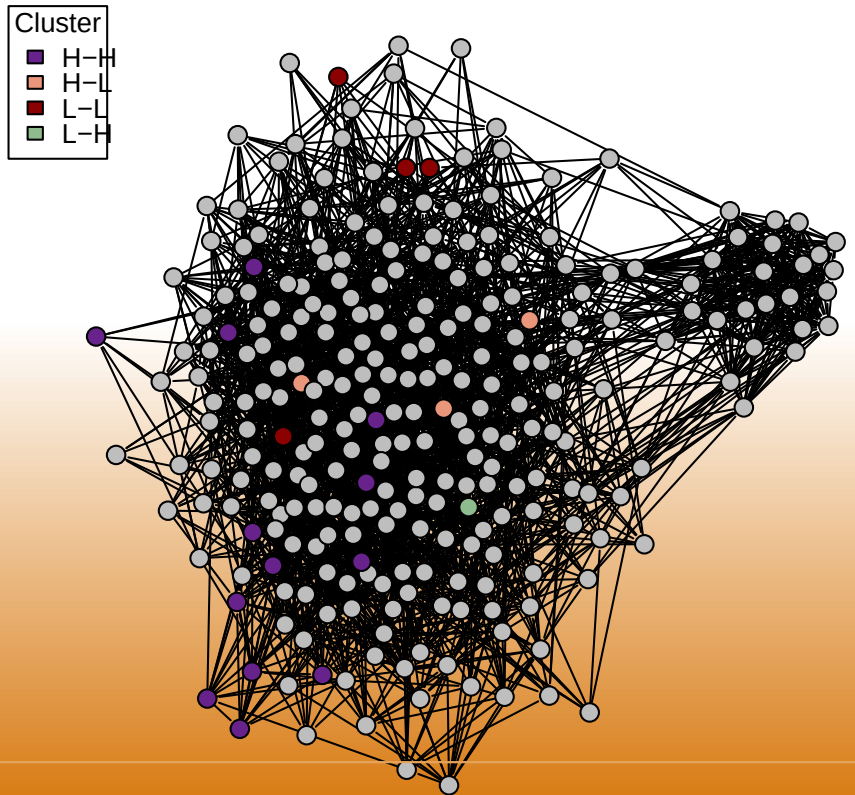
Moran plot : Compare the value of c_i to the mean value of C for the neighbors of i.



Further analysis

Influent points: These genes are candidates to be important genes for the phenotypes of interest because they are in the middle of a high auto-correlation phenomenon.

Influent points



Conclusion and perspectives

Spatial indexes can help to describe and analyze the distribution of a given variable on the nodes of a network.

Conclusion and perspectives

Spatial indexes can help to describe and analyze the distribution of a given variable on the nodes of a network.

Improvements:

- Try to find a way to measure the correlation between the geographical location and the social network (a network where the **nodes are valued by a spatial information**) ;
- Explore other spatial tools for networks: Moran's plot, LISA...

A few references

Boulet, R., Jouve, B., Rossi, F., and Villa, N. (2008).
Batch kernel SOM and related laplacian methods for social network analysis.
Neurocomputing, 71(7-9):1257–1273.

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.

Thank you for your attention...