12010112

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

Thibault Laurent* — **Nathalie Villa-Vialaneix****,*****

* *Toulouse School of Economics*
Université Toulouse 1
Manufacture des tabacs
21 allées de Brienne
31000 Toulouse - France
Thibault.Laurent@univ-tlse1.fr

** *Institut de Mathématiques de Toulouse*
Université Toulouse III (Paul Sabatier)
118 route de Narbonne
31062 Toulouse cedex 9 - France
nathalie.villa@math.univ-toulouse.fr

*** *IUT de Perpignan, Département STID*
Domaine universitaire d'Auriac
Avenue du Dr Suzanne Noël
11000 Carcassonne

ABSTRACT. *A growing number of data are modeled by a graph that can sometimes be weighted: social network, biological network... In many situations, additional informations are provided with these relational data, related to each node of the graph: this can be a membership to a given social group (for social networks) or to a given proteins family (for protein interactions network). In this case, a important question is to understand if the value of this additional variable is influenced by the network. This paper presents exploratory tools to address this question that are based on tests coming from the field of spatial statistic. The use of these tests is illustrated on several examples, all coming from the social network framework.*

RÉSUMÉ. *Un nombre croissant de données sont modélisées sous la forme d'un graphe, éventuellement pondéré : réseaux sociaux, réseaux biologiques... Dans de nombreux exemples, ces données relationnelles peuvent être accompagnées d'une information supplémentaire sur les nœuds du graphe : il peut s'agir de l'appartenance à telle ou telle organisation pour un réseau social ou bien l'appartenance à une famille de protéines pour les réseaux d'interactions de protéines. Dans tous les cas, une question importante est de savoir si cette information supplémentaire est ou non influencée par la structure du réseau. Nous proposons des outils d'exploration de cette*

question, basés sur des tests issus du domaine de la statistique spatiale. L'utilisation de ces tests est illustré au travers de plusieurs exemples, tous issus du domaine des réseaux sociaux.

KEYWORDS: *relational data; social network; Moran's I; join count; permutation test*

MOTS-CLÉS : *données relationnelles ; réseaux sociaux ; I de Moran ; statistiques de comptage ; test de permutations*

1. Introduction

A growing number of real situations are modeled through relational data, i.e., data where the objects under study are not (only) described by informations that fit the standard data analysis framework (real or factor variables) but also by the knowledge of a kind of relationships between the objects. This kind of data include, in particular, social networks, constructed according to a given kind of interactions between persons, or biological networks, where genes or proteins interact to produce a desirable or an unwanted biological phenomenon. The framework of this paper is to deal with relational data that can be modeled by a (possibly weighted) graph such that an additional information is given for each nodes of the graph. This information can be either a factor information or a real value and the underlined problem is to understand if the value observed on the nodes can be influenced by the relations in the network: this question can help to understand the reasons behind the relations in the network or is a prior step before any prediction strategy for nodes that have unknown values.

Among work that deal with additional informations on the nodes of a network are the epidemic propagation models: for example, [NEW 02] deals with the SIR model where differential equations model the spread of a disease's states (susceptible / infective / removed) through a network. These approaches are mostly used for simulation purposes and not for real data analysis. Other approaches involve linear models to explain the spread of a factor information through social relationships: in [CHR 07], the evolution of obesity in a large social network is modelled by a logistic regression having as a covariate the fact that a connected individual is or is not also obese; [VAL 97] models women's contraceptive use of in Cameroon by a diffusion model which is simply a logistic regression taking into account the network auto-correlation effect. In this paper, we concentrate on an exploratory analysis purpose in the case where we do not observe a spread through a network over the time but the status of its nodes at a given moment. We propose the direct use of tools coming from the field of spatial statistic and illustrate it with several examples.

In the following, the relational data are represented by $\mathcal{G} = (V, W)$, a weighted graph with vertices $V = \{x_1, \ldots, x_n\}$ and weights $W = (W_{ij})_{i,j=1,\ldots,n}$ such that $W_{ij} \geq 0$ (and $W_{ij} > 0$ indicates the existence of an edge between nodes x_i and x_j) and $W_{ij} = W_{ji}$ (the weights are symmetric and thus the graph is undirected). Typical examples of such graphs are used to model e.g., social networks (in this case W denotes the number or the intensity of the relation between two persons) but actually, strong analogies exist between relational data and spatial data, the main one being that the geographical proximity between given spatial units is also frequently encoded as a weighted graph.

In addition to the graph, a function $\mathcal{C}$, modelling another information related to the nodes of the graph, is also known:

$$\mathcal{C} : x_i \in V \rightarrow \mathcal{C}(x_i) = c_i.$$

In this paper, c_i is supposed to be a binary information (see Section 2) or a real valued information (see Section 3). Two kind of tests, corresponding to these two cases, are presented in the paper. The use of the tests presented in this paper are illustrated with several examples, relying on a Monte Carlo simulation. The examples are all related to social networks and $\mathcal{C}$ is either the sex of the persons involved in the network (binary information), a geographical location (factor information derived from the binary case) or a date of social activity (real valued information).

2. Case of a binary information

In this section, $\mathcal{C}$ is supposed to take values in $\{0, 1\}$ (without loss of generality, this case models any binary information given on the nodes of the graph).

2.1. *Join count test based on Monte Carlo simulations*

Dealing with data indexed by spatial units ($i \in I$), [MOR 48] introduces a general method to analyze the spatial interaction for a binary variable. More precisely, suppose that $(c_i)_{i\in I}$ is now a binary variable those values are given for spatial units indexed by the finite set I; suppose also that some of these spatial units are joined and other are not. The join count statistic is defined as:

$$JC = \frac{1}{2}\sum_{i\neq j} W_{ij}c_ic_j, \tag{1}$$

in the case where $W_{ij} \in \{0, 1\}$ encodes the fact that the spatial units i and j are joined or not. Then, [CLI 73] extends this measure to arbitrary (and possibly non symmetric) weights able to model more precisely the perception of geographical space; a large literature is devoted to the choice of relevant weights to encode spatial relationships.

This statistic has became very popular as [NOE 70] has proved its asymptotic normality under the assumption of independence of $(c_ic_j)_{ij}$ for distinct pairs of observations. A test for the spatial correlation of $(c_i)_i$ has thus be derived from this result. It relies on the calculus of the mean and standard deviation of the asymptotic law under the null hypothesis and additional assumptions on the sampling distribution.

The same approach can be directly applied to more general networks, in particular social networks, where nodes (e.g., persons involved in the network) play a role similar to spatial units and weights model the intensity of the relations between two nodes, instead of the geographical similarity. Two tests can be derived from the JC index:

– by calculating the index expressed in Equation [1], which is simply:

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

one can test if the number of nodes valued 1 and related to nodes valued in the same way is significantly different (greater or smaller) to what could be expected in the case where no correlation between labels exists in the network;

– by calculating an index similar to Equation [1] but replacing c_i by $\tilde{c}_i = 1 - c_i$, the following index is obtained:

$$JC_0 = \frac{1}{2} \sum_{i \neq j,\, c_i = c_j = 0} W_{ij}.$$ [1]

This index is used to test if the number of nodes valued 0 and related to nodes valued in the same way is significantly different (greater or smaller) to what could be expected in the case where no correlation between labels exists in the network

Unfortunately, these tests are based on the approximation of the distribution of JC by the Gaussian law, which is only valid in an asymptotic way and under other mild conditions. For small networks, this approximation can be bad and a usual method to circumvent this difficulty is to estimate the distribution of JC by a Monte Carlo simulation: the distribution of JC is approximated by the empirical distribution of JC for P permutations of the values of $\mathcal{C}$ among the nodes of the network (where P is large). [CLI 73] show that this approach gives accurate results through several simulation studies.

The following subsections show examples of the application of the approach to a toy social network and a medieval social network. They also provide related comments on the results.

2.2. *Example 1: Sex distribution in "Les Misérables"*

This first example aims at illustrating the use of the join count test for a small social network. This example is described in [KNU 93] and is extracted from the famous French novel "Les Misérables", written by Victor Hugo. From the novel, a weighted graph was defined, counting simultaneous appearances of the 77 characters of the novel in the same chapter. The original data are available at `http://www-personal.umich.edu/~mejn/netdata/lesmis.zip`. A gender information about the character (which is clearly bimodal) was added [2]. The whole graph (network data and additional gender information) is represented in Figure 1. The graph contains 26 (33.8%) women and 51 men. The join count statistic can be used to test four different hypothesis:

1. Note that, similarly, $JC_{0-1} = \sum_{i,\, j:\, c_i=0,\, c_j=1} W_{ij}$ can be used to test the significativity of the proximity between nodes valued with 0 and nodes valued with 1 but these results can be deduced from the other tests as $JC_0 + JC_1 + JC_{0-1} = 2m$ where m is the number of edges of the network.
2. This information is not given with the original data but can be provided on request.

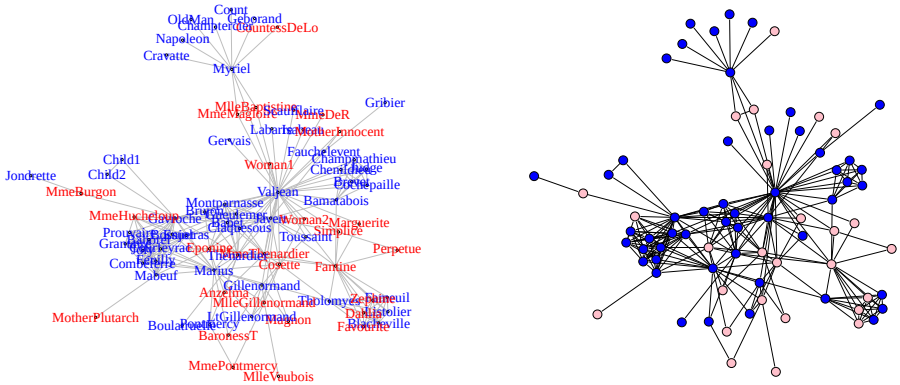

Figure 1. *Co-appearance network from "Les Misérables" with gender information (red or pink: women and blue: men)*

– Is the number of men (M) related to another man significantly greater to what could be expected in the case where no correlation between labels exists in the network?

– Is the number of women (F) related to another woman significantly greater to what could be expected in the case where no correlation between labels exists in the network?

– Is the number of men related to men significantly smaller to what could be expected in the case where no correlation between labels exists in the network?

– Is the number of women related to another woman significantly smaller to what could be expected in the case where no correlation between labels exists in the network?

The R package spdep can be used to compute the test statistics and the p-value based on the comparison with the empirical distribution of JC for P permutations of the values of the genders among the nodes of the network (with the function `joincount.mc`; $P = 1000$ has been used). Figure 2 gives the empirical distribution of joint count JC_F (relations between women). This figure shows that the relations between women in the network tends to be small compared to what is obtained with randomization of the genders among the nodes. Additionally, Table 1 provide the corresponding value for the join count statistics and the p-value associated to the four questions listed above. This table shows that only a relation is significant (with

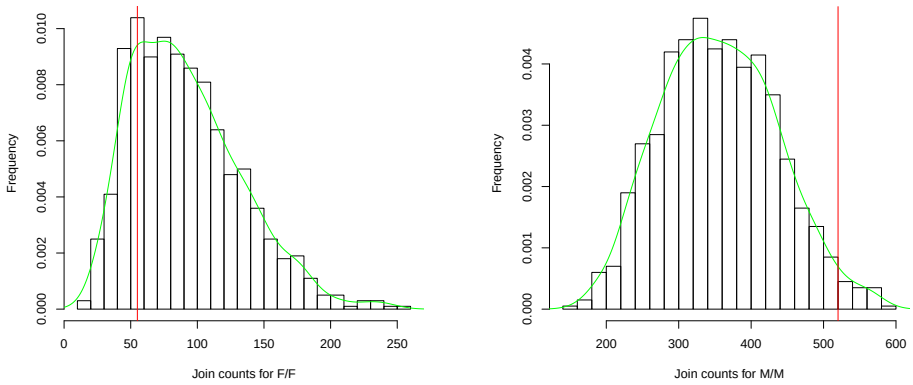

Figure 2. *Empirical distribution and true value (in red) of the join count for the relations "F-F" (left) and "M-M" (right) in the network "Les Misérables"*

Sex	Join count value	Greater	Less
F	55	0.7932 (NS)	0.2068 (NS)
M	520	0.0224 (**)	0.9755 (NS)

Table 1. *Joint count statistics and p-values for the gender relations in the network "Les Misérables". NS means non significant, * means significant at level 10%, ** significant at level 5% and *** significant at level 1%*

a p-value equal to 0.0224): the number of men related to men in the network is significantly greater than what could have been expected in an independent framework. Hence, in the novel of Victor Hugo, not only are the men more numerous but they also tend to interact more often with other men than with women.

2.3. *Example 2: Geographical locations in a medieval social network*

The data used in this example are similar to the data described in [BOU 08] and come from the corpus of documents available at `http://graphcomp.univ-tlse2.fr/` [3]. More precisely, the network has been built from medieval agrarian contracts: the vertices of the network are peasants involved in the contracts and the edges model common quotes in the same contract (the edges are weighted by the number of common quotes). The network is restricted to peasants having at least one activity between 1295 and 1336 (just before the Hundred Years' war). Additionally, the main geographical location of each peasants is also available.

3. Project funded by ANR, number ANR-05-BLAN-0229.

The final graph has 877 vertices and has a density equal to 12.0 %. It is represented by a force directed algorithm (Fruchterman and Reingold as implemented in the R package igraph, see [FRU 91]) in Figure 3 (left); 2 vertices, disconnected from the rest of the graph, were removed from the analysis. 22 geographical locations, all corresponding to villages ("lieu dit" or "paroisse") situated in the seigneury of Castelnau Montratier (Lot) are cited and distributed as in Figure 4. The 5 most frequent locations (Saint-Daunes, Cazillac, Saint-Martin de Valausi, Saint-Julien and Saint-Martin de La Chapelle) are represented on the network in Figure 3 (right).

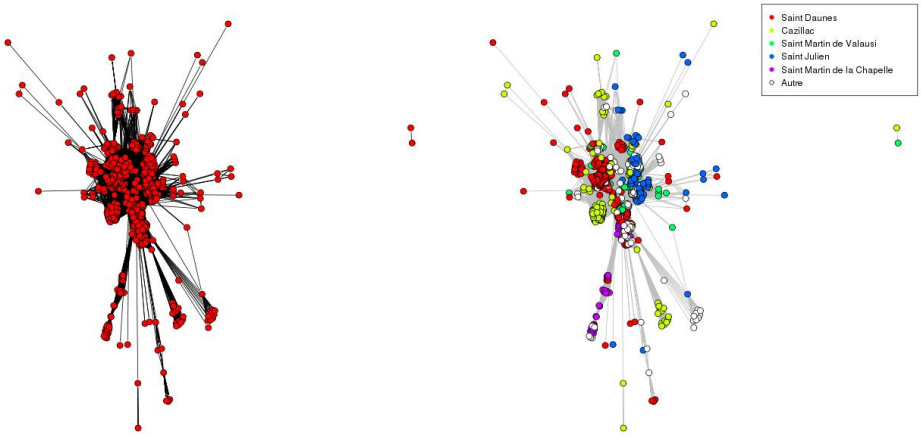

Figure 3. *Medieval social network based on common quotes in agrarian contracts (left) and information about the 5 most frequent geographical locations of the peasants involved in the network (right): see Figure 4 for the distribution of the geographical locations in the network*

From these data, the correlation of each of the 5 most frequent geographical locations among the network has been tested. More precisely, we tested the assumption that the people living in one of those 5 places tend to be more connected (or less connected) to other people living in the same place. To that aim, the join count test has been used with 5 binary variables corresponding to the location in each of the 5 most frequent locations. The results are given in Table 2 for W being the number of contracts between two peasants (weighted graph) and in Table 3 for W being the corresponding binary relation (non weighted graph). The most obvious conclusion is obtained for Saint-Daunes, Cazillac, Saint-Martin de Valausi and Saint-Martin de la Chapelle: for these geographical locations, the number of contracts related to people living in the same village is significantly greater to what could have been expected in the case of a non preferential geographical attachment. For Saint-Julien, the conclusion is a bit harder to understand: the first test, based on the weighted graph (Table 2) shows that peasants living in Saint-Julien tend to interact significantly less often with people having the same geographical location but the test based on the non weighted graph leads to the opposite conclusion. A further analysis helps to explain these dif-

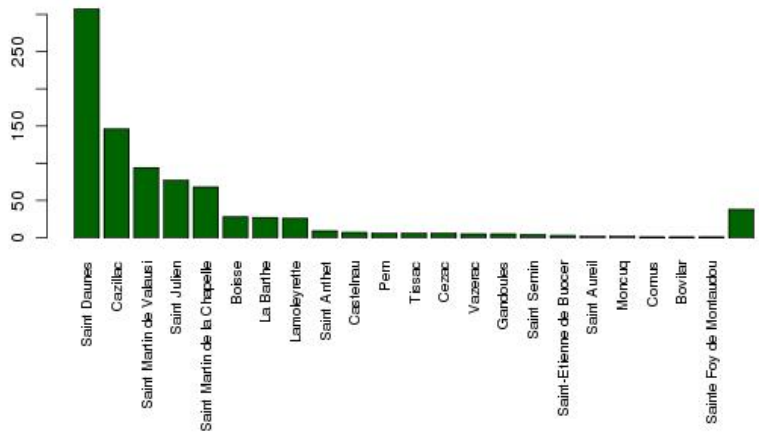

Figure 4. *Distribution of the geographical locations in the medieval social network represented in Figure 3*

Location	Join count value	Greater	Less
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)

Table 2. *Joint count statistics and p-values for the 5 most frequent geographical locations in the weighted medieval social network. NS means non significant, * means significant at level 10%, ** significant at level 5% and *** significant at level 1%*

Location	Join count value	Greater	Less
Saint-Daunes	11 669	0.0010 (***)	0.999 (NS)
Cazillac	2 543	0.0010 (***)	0.999 (NS)
Saint-Martin de Valausi	1 337	0.0010 (***)	0.999 (NS)
Saint-Julien	754	0.0010 (***)	0.999 (NS)
Saint-Martin de la Chapelle	777	0.0010 (***)	0.999 (NS)

Table 3. *Joint count statistics and p-values for the 5 most frequent geographical locations in the non weighted medieval social network. NS means non significant, * means significant at level 10%, ** significant at level 5% and *** significant at level 1%*

ferences: the mean number of contracts for people living in Saint-Julien is very low compared to the other geographical locations (214 whereas the greatest value among the 4 other locations is 1 116 in Saint-Martin La Chapelle and the smallest value is 519 in Cazillac). But the mean number of connected peasants for people living in Saint-Julien is not that different (79 whereas the greatest value is 117 in Saint-Daunes and the smallest value is 67 in Cazillac). Hence, the small value of the join count statistic reported in Table 2 is due to the fact that the peasants in Saint-Julien made only few contracts, even if these contracts are mainly made with people living in the same village (as reported in Table 3).

This simple example illustrates the fact that the use of a weighted or a non weighted graph for the join count statistics can have a strong influence on the result, depending on the question of interest: the number of relationships between peasants living in Saint-Julien is significantly greater to what could be expected but the number of contracts between peasants living in Saint-Julien is significantly smaller to what could be expected because the peasants in Saint-Julien tend to make much less contracts with their relatives than in the other geographical locations.

3. Case of a real valued information

In this section $\mathcal{C}$ takes its values in $\mathbb{R}$.

3.1. *Moran's I and test based on Monte Carlo simulations*

In the spatial statistic framework, the influence of the spatial location on a real valued variable is often accessed through a generalization of the join count statistic of Equation [1]: actually, [MOR 50] introduced the Moran's I statistic which is equal to

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

where $m = \frac{1}{2}\sum_{i\neq j} W_{ij}$ and $\bar{c}_i = c_i - \bar{c}$ with $\bar{c} = \frac{1}{n}\sum_i c_i$. As for the join count statistic, this index has been extended to arbitrary weights by [CLI 73]. Under spatial independence of $\mathcal{C}$, I is also asymptotically distributed as a Gaussian law.

Similarly as for JC, I can be used to test the dependency of the distribution of $\mathcal{C}$ in the network structure. Moreover, the distribution of I in a given network under the independence assumption can be approximated by random permutations of $\mathcal{C}$ among the nodes of the network. As for JC, this lead to the definition of a permutation test which is available through the function `moran.mc` in the R package spdep. Note that in the case of the permutation test, the distribution of I is the same, up to a scaling factor, than those of $\sum_{i\neq j} W_{ij}\bar{c}_i\bar{c}_j$ (as $\frac{2m}{n}\sum_i \bar{c}_i^2$ is constant over all the permutations). This makes this test a direct extension of the permutation join count test presented in Section 2.

As for the join count test, two assumptions can be tested: the first one corresponds to the case where I is significantly greater that the expected value, which is the indication that the value of $\mathcal{C}$ are very similar for nodes that are connected. On the contrary, if I is significantly smaller than the expected value, this means that nodes having very different $\mathcal{C}$ tend to be connected. The following subsection illustrates the first case on the medieval social network introduced in Section 2.3.

3.2. *Example: Date of activity in a medieval social network*

The example used to illustrate the use of this index is the same that the one described in Section 2.3. Here the additional information given for each set is the median of the dates where an activity is reported for the given node (peasant). A simple analysis of this variable is given in Figure 5 where the values of the variables are given for each node of the network (left) as well as the distribution of the median dates in the network (right).

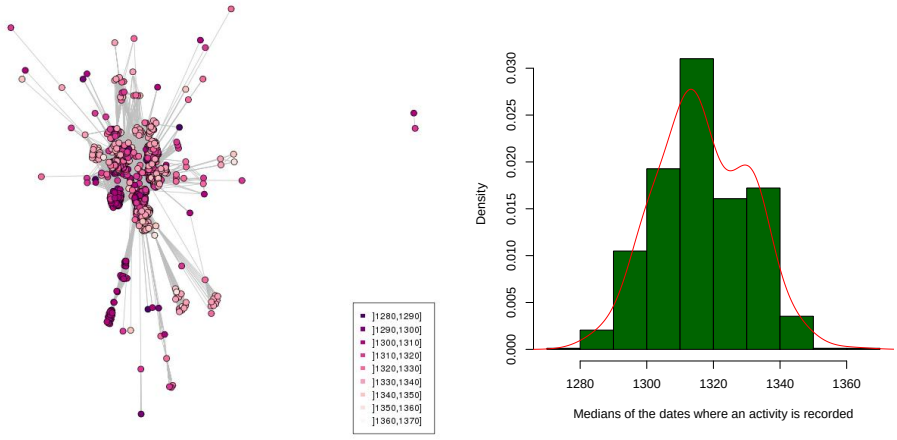

Figure 5. *Representation of the median date for activity (left) and distribution of the median dates (right)*

In this case, the Moran's I based on the weighted graph is equal to $I = 0.2721$ which gives a p-value equal to 0,1% when testing if I is greater to what was expected. The Moran's I based on the non weighted graph is equal to $I = 0.2418$ with the same p-value. Both tests then prove that the peasants in the network tend to be strongly connected only to other peasants having very close median dates of activity. But the whole studied period is only a century long (and that most people have a median date of activity between 1290 and 1340) and the mean length of activity for the peasants in the network is more than 25 years (for peasants having at least two dates reported): this tends to prove that there is a strong generation impact in the way the peasants interact between each others.

4. Conclusion

This paper illustrates how the use of spatial indexes can be useful for exploratory purpose in a network. Based on a similar approach, the links between a network and an additional information given for its nodes could be investigated by using local indexes such as the ones provided by the Moran plot or local Moran indexes (see [ANS 95]). In the same spirit, spatial regression models could provide a way to define prediction models based on a network.

Acknowledgment

The authors are grateful to Florent Hautefeuille, historian in TRACES laboratory, University Toulouse 2 (Le Mirail), for useful informations about the medieval social network and particularly for providing an expert supervision of the definition of the network and for correcting the geographical location information.

5. References

[ANS 95] ANSELIN L., "Local indicators of spatial association-lisa", *Geographical Analysis*, vol. 27, 1995, p. 93-115.

[BOU 08] BOULET R., JOUVE B., ROSSI F., VILLA N., "Batch kernel SOM and related Laplacian methods for social network analysis", *Neurocomputing*, vol. 71, num. 7-9, 2008, p. 1257-1273.

[CHR 07] CHRISTAKIS N., FOWLER J., "The spread of obesity in a large social network over 32 years", *New England Journal of Medicine*, vol. 357, 2007, p. 370-379.

[CLI 73] CLIFF A., ORD J., *Spatial Autocorrelation*, Pion Limited, London, 1973.

[FRU 91] FRUCHTERMAN T., REINGOLD B., "Graph drawing by force-directed placement", *Software-Practice and Experience*, vol. 21, 1991, p. 1129-1164.

[KNU 93] KNUTH D., *The Stanford GraphBase: A Platform for Combinatorial Computing*, Addison-Wesley, Reading, MA, 1993.

[MOR 48] MORAN P., "The interpretation of statistical maps", *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, vol. 10, 1948, p. 243-251.

[MOR 50] MORAN P., "Notes on continuous stochastic phenomena", *Biometrika*, vol. 37, 1950, p. 17-23.

[NEW 02] NEWMAN M., "Spread of epidemic disease on networks", *Physical Review E*, vol. 66, num. 016128, 2002.

[NOE 70] NOETHER G., "A central limit theorem with non-parametric applications", *Annals of Mathematical Statistics*, vol. 41, 1970, p. 1753-1755.

[VAL 97] VALENTE T., WATKINS S., JATO M., VAN DER STRATEN A., TSITSOL L., "Social network associations with contraceptive use among comeroonian women in voluntary associations", *Social Science & Medecine*, vol. 45, 1997, p. 677-687.