Using spatial indexes for labeled network analysis

Thibault Laurent, Nathalie Villa-Vialaneix

Using spatial indexes for labeled network analysis

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 - France

Abstract

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, ou étiquette, 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 la distribution des valeurs de cette étiquette est influencée par la structure même 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 illustrée au travers de plusieurs exemples, tous issus du domaine des réseaux sociaux.

Mots-clés *: données relationnelles, réseaux sociaux, I de Moran, statistiques de comptage, test de permutations, diagramme de Moran, sommets influents*

Abstract

A growing number of data are modeled by a graph that can sometimes be weighted: social network, biological network... In many real world situations, additional information is provided with these relational data, related to each node of the graph. For instance, the nodes of a social network can be labeled by their membership to a social group or, the nodes of a proteins interaction network can be labeled by

proteins families. In this framework, an important question is to understand if the labels of the nodes are, somehow, related to the network topology. To address this question, this paper presents exploratory tools that are based on tests coming from spatial statistics. The use of these tests is illustrated on several examples in the social network framework.

***Key-words**: relational data, social network, Moran's I, join count, permutation test, Moran's plot, influential nodes*

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 information that fits the standard data analysis framework (numerical variables or factors) but also by the knowledge of a kind of relationships between the objects. In particular, these data include social networks, constructed according to a given kind of interactions between persons, or biological networks, where genes or proteins interact to cause a desirable or an unwanted biological consequence. This paper deals with relational data that can be modeled by a (possibly weighted) graph whose nodes are labeled by an additional information. This information can be either a factor or a numerical value and the underlying problem is to understand if the labels are linked to the relations between the nodes in the network: this question can help to understand the reasons underlying the relations in the network or, with an opposite point of view, it is a prior step before any prediction strategy for unlabeled nodes.

Among works that deal with network having labeled nodes, are the epidemic propagation models: for instance, [25] 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 [9], the evolution of obesity in a large social network is modeled by a logistic regression having as a covariate the fact that a connected individual is or is not obese; [29] models women's contraceptive use in Cameroon by a diffusion model which is simply a logistic regression taking into account the network auto-correlation effect. More recently, profile alignment in social networks has raised an increasing interest: analyzing Club Nexus, an online community at Standford University, [1] show an increasing similarity between individuals having a small social distance. From a dynamic point of view, [12, 2] show an increasing similarity between people after their connections in an online community (Wikipedia and aNobii). All these work are based on the simple study of the correlation between the similarity profiles and the distance in the network.

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. An approach is proposed, that combines the topology of the network and the labels of its node in a unique measure, coming from the field of spatial statistics. Spatial statistics deal with geographical entities that are related to each others by a spatial relationship. [13] decomposes the spatial analysis in three main topics depending on the nature of the data: geostatistical data, lattice data and point patterns. Spatial analysis studies spatial entities by using the traditional techniques of statistics and taking into account the topological, geometric or geographic properties of the data. Besides, spatial analysis developed its own methods for detecting and modeling spatial autocorrelation or spatial heterogeneity in the data. One common way to model the proximity between two geographic entities is to define a matrix, W, which contains adjacency information (0/1) or some numerical similarity that is usually a decreasing value of the geographic distance. As pointed out in [15],

> *"Spatial systems and social networks are virtually equivalent phenomena as both can be represented by the adjacency matrix W used to define the spatial or network structure of a system [...]"*

But, until very recently, exploratory tools developed in the field of spatial statistics were not used in social network mining, despite the fact that linear "spatial" models (i.e., models including an auto-correlation part based on the network adjacency matrix) are of a very common use in social network modeling (see [14, 22, 21], among others). A few examples of the use of some of the spatial auto-correlation measures presented in this paper can be found in the literature: in the free statistical software R [28], the package **sna**, dedicated to social network analysis [8], contains a function that calculates autocorrelation indexes, such as Moran's I (see Section 3) and Geary's G. In the marketing discipline, [18] illustrates the use of these two indexes to explain the behavior of a telecommunication company customers. The present article intends to push this application further, not only by illustrating the usefulness of classical spatial autocorrelation measures under various circumstances but also by providing a wider range of tools, including graphical tools such as the Moran's plot and influence measures to highlight important individuals in a social network.

In the following, the relational data are represented by $\mathcal{G} = (V, W)$, a weighted graph with vertexes $V = \{x_1, \ldots, x_n\}$ and weights $W = (W_{ij})_{i,j=1,\ldots,n}$ such that $W_{ij} \geq 0$ (and $W_{ij} > 0$ indicates that there is 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, for instance, social networks (in this case W is the number or the intensity of the relation between two persons).

In addition to the graph, a function $\mathcal{C}$, standing for the labels on the nodes, is also known:

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

In this paper, c_i is supposed to be a binary information (see Section 2) or a numerical information (see Section 3). Two types of tests, corresponding to these two cases, are presented in the paper. The use of the tests are illustrated with several examples, relying on a Monte Carlo simulation which is based on the repetition of the realization of a random process (here the random permutation of the labels among the nodes) to be able to assess its distribution. The examples are all related to social networks.

2 CASE OF BINARY LABELS

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

2.1 Join count test based on Monte Carlo simulations

Dealing with data indexed by spatial units ($i \in I$), [23] introduced a general method to analyze the spatial interaction for a binary variable. More precisely, suppose now that $(c_i)_{i\in I}$ is 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 linked and that others 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 linked ($W_{ij} = 1$) or not ($W_{ij} = 0$). Then, [10] extended this measure to arbitrary (and possibly non symmetric) weights able to model more precisely the perception of the geographical space; a large literature is devoted to the choice of relevant weights to encode spatial relationships.

This statistic became very popular as [27] proved its asymptotic normality (when n tends to infinity) under the assumption of the independence of $(c_ic_j)_{ij}$ for distinct pairs of observations. A test for the spatial correlation of $(c_i)_i$ was 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 would be expected if there was no correlation between labels of linked nodes;

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

$$JC_0 = \frac{1}{2} \sum_{i \neq j,\, c_i = c_j = 0} W_{ij}.$$ [4]

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 would be expected if there was no correlation between labels of linked nodes.

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). This aims at approximating JC distribution under the assumption that the labels are randomly distributed on the network, *provided the network topology and the number of labels of each kind* (contrary to the test based on the asymptotic distribution). Through several simulation studies, [10] showed that this approach gives accurate results.

The following subsections respectively illustrate the use of this index on a small social network and on a medieval social network.

2.2 Example 1: Gender distribution in "Les Misérables"

This first example aims at illustrating the use of the join count test for a small social network used as a simple example. This example is described in [20] and is extracted from the famous French novel "Les Misérables", written by Victor Hugo. From the novel, a weighted graph was built, counting simultaneous appearances of the 77 characters of the novel in the same chapter [5]. A gender information (which is clearly bimodal), about the characters, was added. This information is not given with the original data but can be found,

4. Note that, similarly, $JC_{0-1} = \sum_{i,\, j:\, c_i=0,\, c_j=1} W_{ij}$ can be used to test the significance 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.

5. The original data are available at `http://www-personal.umich.edu/~mejn/netdata/lesmis.zip`

as well as the code used to perform the simulations related to this network, in supplementary material available at `http://www.nathalievilla.org/suppmaterial-I3/`. The whole graph (network data and additional gender information) is displayed in Figure 1. The graph contains 26

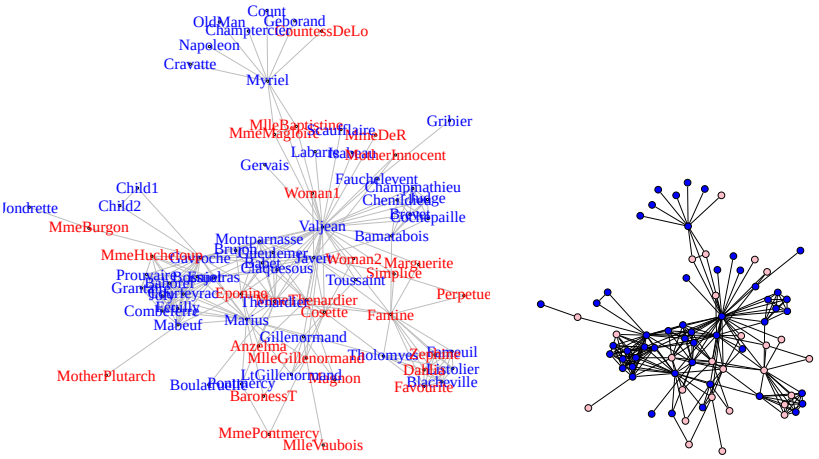

Figure 1: Co-appearance network from “Les Misérables” with gender labels (red or pink: women and blue: men).

(33.8%) women and 51 men. The join count statistic can be used to test four different assumptions:

- Is the number of men (M) related to another man significantly greater to what would be expected if labels of connected nodes were not correlated?
- Is the number of women (F) related to another woman significantly greater to what would be expected if labels of connected nodes were not correlated?
- Is the number of men related to men significantly smaller to what would be expected if labels of connected nodes were not correlated?
- Is the number of women related to another woman significantly smaller to what would be expected if labels of connected nodes were not connected?

The R package **spdep** [6] can be used to compute the test statistic and the p-values (i.e., the probability to obtain the observed value of the statistic under the null assumption) 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`; $P = 1\ 000$ was used). Figure 2 gives the empirical distribution of join count JC_F and JC_M (respectively, relations between women and relations between men). This figure shows that the number of relations between women in the network

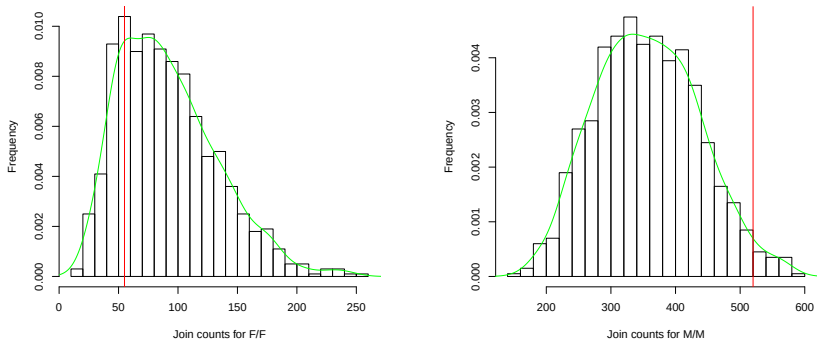

Figure 2: Empirical distributions and true values (in red) of the join count for the relations "F-F" (left) and "M-M" (right) in the network "Les Misérables".

tends to be small compared to what is obtained by randomizing the genders among the nodes whereas the number of relations between men tends to be large. Additionally, Table 1 provides the corresponding values for the join count statistic and the p-values associated to the four questions listed above. This table shows that only a relation is significant (with a p-value equal to

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

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

0.0224): the number of men related to another man in the network is significantly greater than what was 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 [7] and come from the corpus of documents available at `http://graphcomp.univ-tlse2.fr`[6]. More precisely, the network was built from medieval

6. Project "Graph-Comp" funded by the ANR, number ANR-05-BLAN-0229.

agrarian contracts: the vertexes 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 peasant is also available.

The final graph has 877 vertexes and has a density equal to 12.0 %. It is displayed in Figure 3 (left) by using a force directed algorithm (Fruchterman and Reingold as implemented in the R package **igraph**, see [16]). 2 vertexes, that are 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, a French "département" in the South of France) are cited and distributed as in Figure 4. The 5 most frequently cited locations (Saint-Daunes, Cazillac, Saint-Martin de Valausi, Saint-Julien and Saint-Martin de La Chapelle) are displayed on the network in Figure 3 (right).

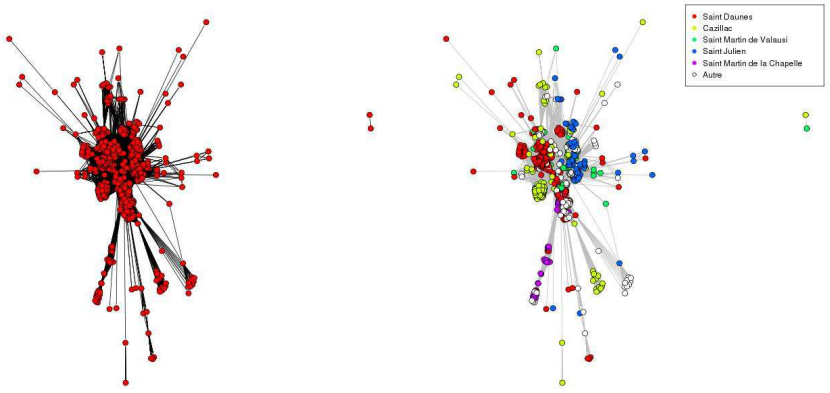

Figure 3: Medieval social network based on common quotes in agrarian contracts (left) and information about the geographical locations of the peasants involved in the network (right). Only the 5 most frequently cited geographical locations are displayed: see Figure 4 for the distribution of all geographical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

For these data, the correlation of each of the 5 most frequent geographical locations to the network topology was 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 was used with 5 binary variables corresponding to the location in each of the 5 most frequently cited places. 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, i.e., only the fact that two peasants have made

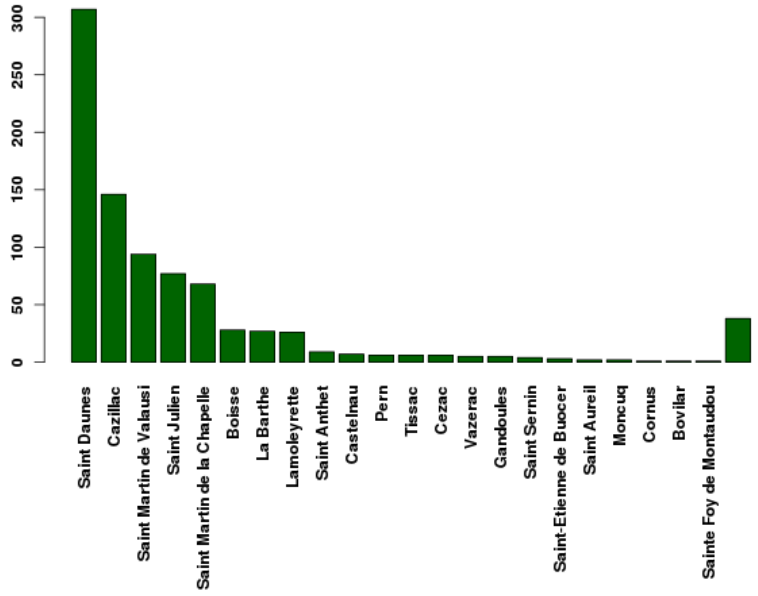

Figure 4: Distribution of the geographical locations in the medieval social network represented in Figure 3.

at least a common contract together is used). The most obvious conclusion

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: Join count statistic and p-values for the 5 most frequently cited geographical locations in the weighted medieval social network. NS means non significant, * means significant at a level of 10%, ** significant at a level of 5% and *** significant at a level of 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: Join count statistic and p-values for the 5 most frequently cited geographical locations in the non weighted medieval social network. NS means non significant, * means significant at a level of 10%, ** significant at a level of 5% and *** significant at a level of 1%

is obtained for Saint-Daunes, Cazillac, Saint-Martin de Valausi and Saint-Martin de la Chapelle: for these places, the number of contracts related to people living in the same village is significantly larger than what was expected in the case where the peasants living in these villages would have no special tendency to interact with people living inside or outside their own village. For Saint-Julien, the conclusion is a bit harder to understand: the first test, based on the weighted graph (Table 2), shows that the peasants living in Saint-Julien tended 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 this difference: the JC value obtained for Saint-Julien in Table 3 means that 754 couples of peasants living in this town made at least one transaction. But, the JC value obtained for Saint-Julien in Table 2 indicates that the total number of transactions between these couples is equal to 1 172, i.e. 1.5 contracts per couple on average. This value is very low compared to the other ratios (15 in Saint-Martin de Valausi, 13 in Saint-Martin de la Chapelle, 9.5 in Saint-Daunes and 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 were 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 statistic can have a strong impact on the result, depending on the question under focus: the number of connections between peasants living in Saint-Julien is significantly greater than what was expected but the number of contracts between peasants living in Saint-Julien is significantly smaller to what was expected because the peasants in Saint-Julien tended to make much less contracts with the people they were connected to, than the peasants living in the other villages did.

3 CASE OF NUMERICAL LABELS

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

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

In the spatial statistics framework, the influence of the spatial location on a numerical variable is often assessed through a generalization of the join count statistic of Equation (1). Indeed, [24] 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 was extended to arbitrary weights by [10]. Under spatial independence of $\mathcal{C}$, I is also asymptotically distributed as a Gaussian random variable.

Similarly as JC, I can be used to test the correlation between labels of connected nodes. Moreover, under the assumption of the independence between labels and connections, the distribution of I can be approximated by random permutations of $\mathcal{C}$ among the nodes of the network. This leads 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 means that, for connected nodes, the values of $\mathcal{C}$ are very similar and much larger or lower than the mean. On the contrary, if I is significantly

smaller than the expected value, nodes having strong and opposite values of $\mathcal{C}$ compared to the mean, tend to be connected.

3.2 Moran's plot and influential nodes

Another very common tool to analyze spatial auto-correlation of numerical variables is the Moran's plot (see [3]): it displays, for the variable under study, the original value on the horizontal axis and the sum of values observed among the neighbors (according to the adjacency matrix W) on the vertical axis. This last quantity is called the *lag*. If W is row normalized (i.e., each row of W is scaled such that it sums to 1; see Section 3.4 for an application), the vertical axis displays the average value of the variable over the neighbors of a node, according to the weights W. In the case of a centered variable, Moran's I is exactly the slope of the linear trend of the Moran's plot. Moran's plot is usually divided into four quadrants that correspond to different types of spatial correlation. Spatial clusters in the upper right (High-High) and lower left (Low-Low) quadrants, and spatial outliers in the lower right (High-Low) and upper left (Low-High) quadrants:

- quadrant H-H contains nodes, x_i, for which c_i is above the mean $\bar{c}$ and for which the average (or the sum if W is not row-normalized) value of $(c_j)_j$, for $(x_j)_j$ connected to x_i, is also above the mean;
- quadrant H-L contains nodes, x_i, for which c_i is above the mean $\bar{c}$ but for which the average value of $(c_j)_j$, for $(x_j)_j$ connected to x_i, is below the mean;
- quadrant L-H contains nodes, x_i, for which c_i is below the mean $\bar{c}$ but for which the average (or the sum if W is not row-normalized) value of $(c_j)_j$, for $(x_j)_j$ connected to x_i, is above the mean;
- finally, quadrant L-L contains nodes, x_i, for which c_i is below the mean $\bar{c}$ and for which the average (or the sum if W is not row-normalized) value of $(c_j)_j$, for $(x_j)_j$ connected to x_i, is also below the mean.

In addition, as Moran's plot stands in the linear model framework, influence measures are often calculated: these identify the nodes that have a strong influence on several indexes, for example the value of the slope of the linear trend (those that would lead to a large change in the value of the slope if removed from the dataset). Unlike node degrees, influence measures are global network measures: they take into account the whole distribution of the labels values all over the network. Moreover, the function `influence.measures` calculates 6 influence measures: the impacts of a single observation on each regression coefficient is shown by the DFBETA (`dfb.1` for the y-intercept and `dfb.x` for the slope). The COVRATIO measure (`cov.r`) estimates the effect of the observation on the efficiency of the estimation process in the linear model. The Cook's distance (`cook.d`) captures the impact of an observation from two sources: the size of changes in the predicted values when the case is omitted (outlying studentized residuals) as well as the observation's distance from the other observations (leverage).

The DFFIT measure (`dffit`) indicates how much the fitted values change when the case is deleted. The diagonal of the hat matrix (`hat`) is a measure of the distance of the observation from the mean center of all observations; large values also indicate that the observation is disproportionately responsible (compared to the other nodes) for the prediction of the dependent variable value (see [5, 11, 19] for further details about these measures).

Moreover, the function provides a facility to identify cases that are influential with respect to any of these measures: as shown in the following examples, these nodes may have a special behavior in the network and their identification can provide useful directions for the interpretation. The following subsections provide examples of the use of Moran's I as well as illustrations of the way the Moran's plot can be interpreted in the social network framework.

3.3 Example 1: dates in a medieval social network

The first example relies on the network that was described in Section 2.3. Here the additional information given for each node is the median of the dates where an activity (i.e., a citation in a contract) is reported for the given node (i.e., a peasant). A simple analysis of this variable is given in Figure 5 where the median dates are given for each node of the network (left) as well as the histogram of the median dates in the network (right). Because peas-

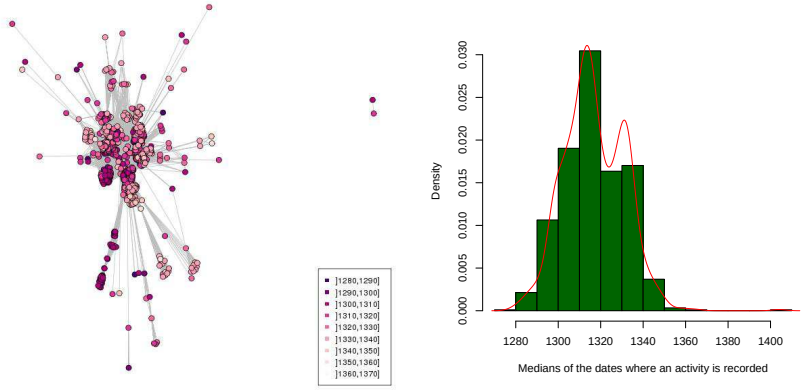

Figure 5: Representation of the median date for activity (left) and distribution of the median dates (right).

ants having an activity centered around 1320 are more likely than others to make contracts during the period (by definition), they have a larger number of connections. Thus, to avoid this size effect, we used a row normalized

network, i.e., we used the network weighted by:

$$\widetilde{W}_{ij} = \frac{W_{ij}}{\sum_{k=1}^{n} W_{ik}}$$

where W_{ij} is equal to 1 if nodes i and j are linked and to 0 otherwise (hence, W is the adjacency matrix of the unweighted graph described in Section 2.3). Note that $\widetilde{W}$, which is of a common use in spatial statistics, is not symmetric.

As the set of all permutations of the dates among the nodes of the network is much larger than the set of all permutations of a binary variable, we used $P = 10\ 000$ random permutations for the Monte Carlo simulation (instead of 1 000 in Section 2.3). In this case, Moran's I based on the unweighted graph is equal to $I = 0.4232$ whereas the largest value for the Moran's I calculated over the 10 000 permutations is only equal to 0.0342. This means that the peasants in the network tend to be strongly connected to peasants having very close median dates of activity. But the whole studied period is only a century long and most people have a median date of activity between 1290 and 1340. Moreover, the average length of activity for the peasants in the network is more than 25 years (for peasants having at least two dates reported): this could mean that there is a strong generation impact in the way the peasants interact between each others.

Figure 6 (left) gives the Moran's plot which exhibits a good linear trend and helps to emphasize characters (influential nodes) having a tendency to interact mainly with people having an earlier (or a later) median date of activity than the one expected. In Figure 6 (right), 67 peasants are identified

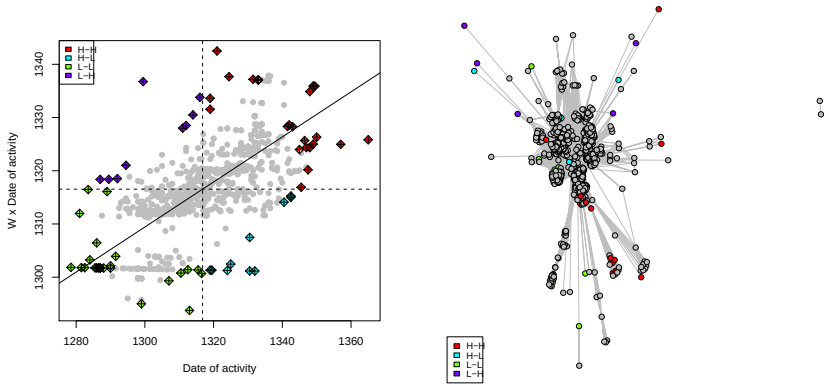

Figure 6: Moran's plots for the median dates of activity (left) and influential nodes on the medieval network (right).

as influential nodes and are displayed with a color corresponding to their quadrant. As the database contains some errors related to a large number of namesakes, these nodes are good candidates for a further study and check of

the validity of the information recorded (especially those of quadrants H-L and L-H). Such an automatic procedure is very useful in that context.

3.4 Example 2: relating the number of connections of the nodes to the network structure

In this section, we first illustrate the use of Moran's I and of the Moran's plot with the same simple example as in Section 2.2 [7] and with an additional (larger) real social network but in a different framework than in the previous section: the labels of the nodes are their degrees and are thus related to the network structure. The underlying issue is then to answer the following question:

Do nodes having many connections have a tendency to be connected to each others?

This is a well known fact that high degree nodes tend to attract relations [4]. But, knowing this structural property, do nodes having a high degree tend to be related to each others more than they should? As pointed out by an anonymous reviewer, in this situation, Monte Carlo simulations should not be performed by permuting values of the degrees of the nodes because such a process does not respect the structural relation between the degrees and the network topology. The natural dependency of the labels (degrees) and the network topology results in a larger Moran's I value than with any other label not related to the network topology: there is indeed a natural network auto-correlation for degrees because nodes having a large degree are easier to be connected with [8]. Hence, Monte Carlo simulations should be performed by randomly permuting edges (instead of labels) *while keeping the degree of each node*. Indeed, this provides the Moran's I distribution under the null model where the degree distribution is known but the edges in the network are displayed totally at random according to that distribution. This is the same null model than the one used to compute the modularity criterion for clustering the vertexes of the graph [26]. Compared to the previous case where there was no relation between the network topology and the labels of the nodes, the empirical distribution does not aim at approximating the same null assumption:

– when permuting nodes labels, the null assumption is that of a random distribution of the labels among the nodes *given the network topology*. Hence, this makes no sense when the labels actually depend on the network topology itself;

7. The code used to perform the analysis of this network is available as supplementary material at `http://www.nathalievilla.org/suppmaterial-I3`.

8. note that expected values for Moran's I could be calculated by a direct analytical approach for random graph models where the dependency structure between weights is known but this issue is out of the scope of this paper

– when permuting edges while keeping the same degrees, the null assumption is that of a random distribution of the links *given the degree distribution*. This case is unusual in spatial statistics because links are direct consequences of the geographical properties of the spatial entities.

This question is first illustrated by the non-weighted symmetric graph deduced from the co-appearance network in "Les Misérables". More precisely, if W is the adjacency matrix of the graph described in Section 2.2, the graph used in this section is the one having for adjacency matrix $\widetilde{W}$ where $\widetilde{W}_{ij} = 1$ if $W_{ij} > 0$ and $\widetilde{W}_{ij} = 0$ otherwise. The numerical variable of interest is simply the number of persons known by each character of the novel (hence, the degree of the node). The underlying question addressed by Moran's I is then:

Do people having many connections in the novel have a tendency to be connected to each others?

Figure 7 (left) displays the number of connections for each character in the co-appearance network; most nodes have a small number of connections (one or two) whereas a few numbers of characters have a much larger number of connections: Valjean (who is the main character of the novel), Gavroche and Marius (see Figure 1 for the complete correspondence between the nodes and the names). In this example, the Moran's I is equal to 0.5148 which is

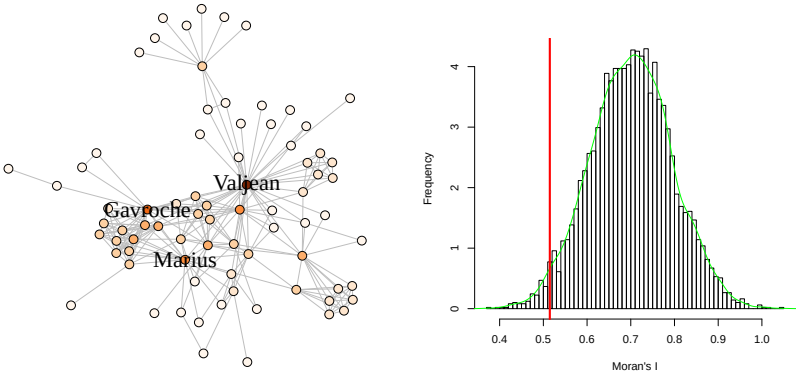

Figure 7: Left: Number of connections for each character in the co-appearance network from "Les Misérables" (dark colors correspond to characters having many connections). Right: Histogram of the 5 000 I values calculated over the Monte Carlo simulation and true I (red vertical line). The names of the characters are given in Figure 1.

significantly smaller than the expected value under the null assumption (the estimated p-value is equal to 2.2%). This dependency is illustrated by the histogram given in Figure 7 (right): it shows that the real I is smaller than most

of those found by randomly permuting the edges of the network 5 000 times and leads to the conclusion that the main characters tend to mainly interact with less important characters. This result could be a consequence of the star shaped structure observed around important characters in the network.

Moreover, Figure 8 provides the Moran's plot for the number of connections in the graph. This figure exhibits a very good fit between the number of connections and its lag (the total number of connections of the neighbors for a given node), except mainly for Valjean whose lag is much smaller than what was expected considering its degree (but remember that Valjean is the most important character of the novel). In this plot, nodes having a large influence for at least one of the 6 influence measures performed, are emphasized: three of them are central characters with large degrees: Valjean (influential for all the 6 measures), Gavroche (influential for COVRATIO and hat measures) and Javert (influential for COVRATIO). But the last one, Myriel (influential for COVRATIO) is a much interesting case: he is an influential node because the total number of connections of his neighbors is much smaller than what was expected from the linear trend between the number of connections and the lag which makes it an influential node because it has a substantial influence on the set of coefficients (both the y-intercept and the slope). Myriel is a character surrounded by many unimportant characters and can be seen as an outlier. This character wouldn't have been highlighted by usual network indexes (such as, e.g., degrees or betweenness) but is nevertheless an important character in the novel (even if not one of the most important): he plays a key role because he is the compassionate country priest whose generosity is the origin of Valjean's redemption. This small and

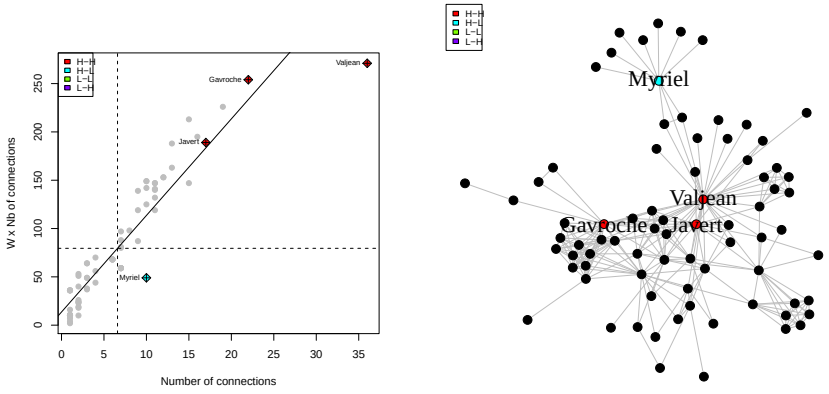

Figure 8: Moran's plot of the number of connections in the co-appearance network in "Les Misérables" (left) and corresponding influential nodes emphasized on the network (right). The names of the characters are given in Figure 1.

simple example illustrates how the influence measures can help to identify important nodes and/or outliers that couldn't have been identified obviously otherwise.

A similar issue is addressed by the study of a larger network representing the email exchanges between members of the University Rovira i Virgili (Tarragona), presented in [17] [9]. The unweighted and symmetric graph coming from that network is used in the simulations: it is connected and contains 1 133 nodes. Using the Moran's I, the following assumption is tested:

Do people with a large number of contacts have a tendency to be connected to each others?

The numerical variable tackled in this study is the the number of people with which a given member of the University exchanges emails; hence, it is the degree of the node.

As the number of nodes in this network is larger than in the previous simple example, 10 000 random permutations were used in the Monte Carlo simulation. Among all these permutations, Moran's I was never greater than 1.022 whereas the observed Moran's I is equal to 1.081. Contrary to the previous example, this leads to the conclusion that people with a large number of contacts have a strong tendency to be connected to each others. It would be an interesting issue to check if the difference between the network coming from "Les Misérables" and this real network is representative of a narrative network structure versus a real world social network structure.

In addition, Figure 9 provides the Moran's plot of the number of contacts over the network and emphasizes the influential nodes on the emails exchange network. As in the previous simple example, the correlation between the number of contacts and the lag is strong and several influential nodes are identified. Some of them are nodes with a large number of exchanged emails whereas others exhibit an unusual behavior, having a lag value (the total number of contacts of the neighbors) either larger or smaller than what was expected. 93 influential nodes are detected this way, providing tips to identify important people in the network.

4 Conclusion

This paper illustrates how the use of spatial indexes can be useful for exploratory purpose in a network framework. More precisely, the distribution of a given variable, that can be either a numerical variable or a factor, can be related to the network structure, in a significant way, by a simple and fast Monte Carlo approach. Moreover, the use of a linear model between a numerical variable and its lag can help to visualize the strength of this dependency and to identify influential nodes. However, one of the main drawback

9. available at `http://deim.urv.cat/~aarenas/data/xarxes/email.zip`

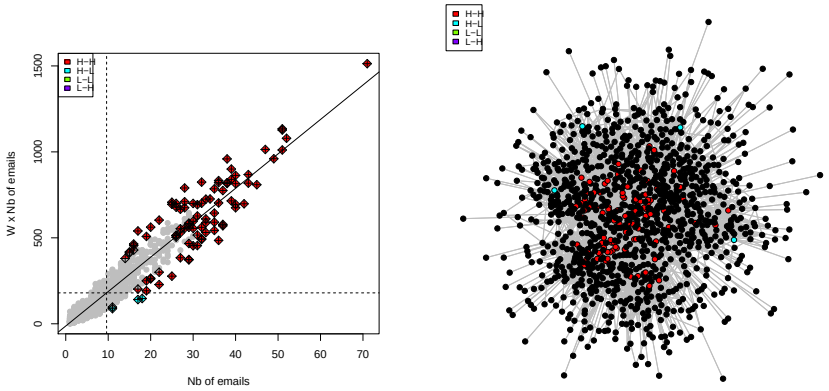

Figure 9: Moran's plot of the number of contacts in the email network of the University Rovira i Virgili (left) and corresponding influential nodes on the network (right).

of this approach is that it is more suited to the case of a numerical labeling. When the labels of the nodes are factors, the join count statistic can be useful for binary labels or for labels having a small number of possible values (by creating separate binary problems) but it is unworkable for tags or keywords analysis (on web pages).

Further studies would lead to study local indexes (such as the local Moran index; see [3]) or to clarify the relevance of the use of the weighted or unweighted adjacency matrix (see Section 2.3) of the network or of its row normalized version (as in Section 3.3), in various applications. Finally, once the dependency between the labels of the nodes and the network structure is established, spatial linear models would be a relevant approach to define predictive models able to infer the labels of unlabeled nodes.

ACKNOWLEDGMENT

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

The authors thank the anonymous reviewers for their valuable comments and suggestions that helped to improve the quality of the paper.

REFERENCES

[1] L.A. Adamic, O. Buyukkokten et E. Adar. A social network caught in the web. *First Monday*, 8, 2003.

[2] L. Aiello, A. Barrat, C. Cattuto, G. Ruffo et R. Schifanella. Link creation and profile alignment in the aNobii social network. In *Proceedings of the Second IEEE International Conference on Social Computing (SocialCom)*, Minneapolis, USA, August 20-22 2010.

[3] L. Anselin. Local indicators of spatial association-lisa. *Geographical Analysis*, 27:93–115, 1995.

[4] A. Barabási et R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[5] D.A. Belsley, E. Kuh et R.E. Welsch. *Regression Diagnostics*. Wiley, New York, 1980.

[6] R.S. Bivand, E.J. Pebesma et V. Gomez-Rubio. *Applied Spatial Data Analysis with R*. Springer, New York, USA, 2008.

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

[8] C.T. Butts. Social network analysis with sna. 24(6):1–51, 5 2008.

[9] N.A. Christakis et J.H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357:370–379, 2007.

[10] A.D. Cliff et J.K. Ord. *Spatial Autocorrelation*. Pion Limited, London, 1973.

[11] R.D. Cook et S. Weisberg. *Residuals and Influence in Regression*. Chapman and Hall, London, 1982.

[12] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg et S. Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th SIGKDD*, pages 160–168, 2008.

[13] N.A.C. Cressie. *Statistics for Spatial Data*. Wiley-Interscience, 1993.

[14] P. Doreian, K. Teuter et C.H. Wang. Network auto-correlation models - some Monte-Carlo results. *Sociological Methods and Research*, 13:155–200, 1984.

[15] S. Farber, A. Pàez et E. Volz. Topology and dependency tests in spatial and network auto-regressive models. *Geographical Analysis*, 41(2):158–180, 2009.

[16] T. Fruchterman et B. Reingold. Graph drawing by force-directed placement. *Software-Practice and Experience*, 21:1129–1164, 1991.

[17] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt et A. Arenas. Self-similar community structure in a network of human interactions. *Physical Review E*, 68(065103(R)), 2003.

[18] M. Haenlein. A social network analysis of customer-level revenue distribution. *Marketing Letters*, 22(1):15–29, March 2011.

[19] J.F. Hair, R.L. Tatham, R.E. Anderson et W. Black. *Multivariate Data Analysis (5th Edition)*. Prentice Hall, March 1998.

[20] D.E. Knuth. *The Stanford GraphBase: A Platform for Combinatorial Computing*. Addison-Wesley, Reading, MA, 1993.

[21] R.T.A. Leenders. Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks*, 24:21–47, 2002.

[22] P.V. Marsden et N.E. Friedkin. Network studies of social influence. In S. Wasserman et J. Galaskiewicz, éditeurs, *Advances in Social Network Analysis: Research in the Social and Behavioral Sciences, SAGE*, pages 3–25, Thousand Oaks, 1994.

[23] P.A.P. Moran. The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 10:243–251, 1948.

[24] P.A.P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37:17–23, 1950.

[25] M.E.J. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(016128), 2002.

[26] M.E.J. Newman et M. Girvan. Finding and evaluating community structure in networks. *Physical Review, E*, 69:026113, 2004.

[27] G.E. Noether. A central limit theorem with non-parametric applications. *Annals of Mathematical Statistics*, 41:1753–1755, 1970.

[28] R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2010. ISBN 3-900051-07-0.

[29] T.W. Valente, S.C. Watkins, M.N. Jato, A. van der Straten et L.P.M. Tsitsol. Social network associations with contraceptive use among comeroonian women in voluntary associations. *Social Science & Medecine*, 45:677–687, 1997.