Spatial correlation in bipartite networks: the impact of the geographical distances on the relations in a corpus of medieval transactions

Nathalie Villa-Vialaneix*, Bertrand Jouve**
Fabrice Rossi*, Florent Hautefeuille***

*SAMM, UniversitÃl' Paris 1, 90 rue de Tolbiac, 75013 Paris - France
{nathalie.villa,fabrice.rossi}@univ-paris1.fr,
http://www.nathalievilla.org, http://www.apiacoa.org
**Laboratoire ERIC-IXXI, UniversitÃl' LumiÃÍre Lyon 2
5 avenue Pierre MendÃÍs-France, 69676 Bron cedex - France
bertrand.jouve@univ-lyon2.fr
http://eric.univ-lyon2.fr/∼bjouve/BER_FR.html
***TRACES, UniversitÃl' Toulouse 2
5, allÃl'es Antonio Machado, 31058 Toulouse Cedex 9 - France
florent.hautefeuille@univ-tlse2.fr
http://w3.terrae.univ-tlse2.fr/spip/spip.php?article35

Abstract. In this article, the influence between a spatial information and interactions between individuals is addressed. This issue is illustrated through the analysis of a corpus of notarial acts established during the Middle Ages. In this corpus, the persons interact in common transactions that are geolocalized. The present work tries to quantify the impact of this spatial information on the relations between people. As the spatial information as well as the relations between individuals are derived from the same source (the transactions in which the individuals have been involved), a standard Mantel test (Mantel, 1967) is not suited to address this issue. A similar methodology, based on the adaptation of the original permutation test, is thus proposed and illustrated in that context. [1]

1 Introduction

Relational models are increasingly used as a theoretical framework in a number of real life situations. Indeed, "graphs" are useful models to mathematically describe co-expression networks and metabolic pathways (biology), social relationships, computer networks or any other kind of data where the objects under study, the "entities", are related to each others by a given type of relation (Dorogovtsev and Mendes, 2003; Newman et al., 2006). In the simplest case, the graph is made of N nodes (or vertices) $V = \{x_1, \ldots, x_N\}$, each one representing an entity, and a set of edges $E \subset V \times V$, indicating if a pair of entities is linked by a relation.

1. This work is partially supported by ANR ModULand (Programme blanc SHS 1 2011 0).

Additionally, information can be added to the edges (weights, orientation, type...) or to the nodes (entity description).

This paper deals with the case where the nodes of the graph are described by a spatial information. In this case, one may be interested to understand if the relations described by the graph are somehow linked to the geographical locations of the nodes. In (Laurent and Villa-Vialaneix, 2011), this question is tackled by the use of a Join Count (JC) permutation test (Moran, 1948). The method is tested on a graph built from a large corpus of medieval notarial acts: the nodes of the graph are the individuals actively involved in the transactions and the edges of the graph model a common citation of two individuals in the same transaction. Additionally, each node (individual) is labeled by the most common place where he is cited. Then, different JC permutation tests for the five most common locations are computed. This approach shows a correlation between relations in the network and the geographical labels, for the five most important villages: people located in a given village tend to interact significantly more often with people located in the same village than with the others.

But JC statistics can only measure the correlation of the graph structure with a binary variable, leading to two main simplifications of the problem: each node (tenant in this case) has to be related to a single location and the correlation between the network structure and each location has to be tested independently. Actually, people are often localized in more than one place (with different citation frequencies in the different places) and the use of the full geographical information attached to each individual might provide additional insights on the link between the network structure and the geographical locations of the nodes. Moreover, the individuals' locations are not simply factors and it might also be interesting to take into account the distances between villages while evaluating the correlation between social relations and the spatial information. Such a study has been started in (Miquel, 2011) and is improved in this paper. In particular, a new permutation based test is introduced to overcome the fact that the spatial information and the relations between individuals come from the same source. This methodology can be viewed as a generalization of the Mantel test to distance matrices that are linked by a common relational model. This approach can cope with multiple spatial information per node (by computing an average spatial location) and really relates the geographical topology to the network structure.

The rest of the paper is organized as follows: Section 2 describes the data set, the network model and the tested assumptions. Section 3 describes how Mantel test has been adapted to that framework and presents three different ways to address the question. Finally, the results are given and discussed in Section 4.

2 Description of the data set

Whereas the methodology described in the next section is general enough to deal with a variety of data, this article focuses on a graph built from a corpus of medieval notarial acts. This corpus has already been described in (Boulet et al., 2008; Rossi and Villa-Vialaneix, 2011; Rossi et al., 2012; Hautefeuille and Jouve, 2012) and the relational model on which the present paper is based is described in (Rossi et al., 2012). In short, the network intends to model a corpus of documents preserved at the "Archives dÃl'partementales du Lot" (Cahors, France). The corpus is the original work of a feudist who was hired to collect (and re-write) all the notarial acts that mentioned rents and were established for the sake of one of the lord of the

"seigneurie" named "Castelnau Montratier". Hence, all documents in the corpus share one main common characteristic: the transactions are all related to the same "seigneurie", Castelnau Montratier, located in the present-day French dÃl'partement Lot (South West of France). The "seigneurie" contained about 40 villages (parishes), for a total area of approximately 300 km^2. The original acts were established between 1238 and 1768, with a shifting abundance depending on the time period. The whole corpus can be seen as a very representative sample of the land charters established in this "seigneurie" during that period. Each act contains one or several transactions that have been recorded in a large database freely available on line at `http://graphcomp.univ-tlse2.fr`. More than 75% of the whole corpus has already been digitized with a priority put on an homogeneous geographical capture and on transactions established before 1500.

Each transaction contains various information, with different degrees of precision. Generally, the transaction comes with a date, the names of the lords and tenants directly involved (as active participants) in the transaction, the location of the land that is concerned by the transaction, the names of the neighbor(s) of the land concerned by the transaction and the name of the notary who wrote the transaction out. Other information, such as the status of the cited individuals, is also sometimes available. Then, a relational model is derived from the database where all these informations have been saved:

— the nodes model both the transactions and the individuals directly involved in those transactions (hence notaries and neighbors are excluded);
— two nodes are linked by an edge if the individual represented by one node is directly involved in the transaction represented by the other node;
— the nodes representing transactions are labeled with the name of the parish where the land concerned by the transaction is located, as it is reported in the transaction.

This relational model is a *bipartite graph*: in bipartite graphs, the nodes can be divided into two distinct groups. Nodes in the same group are never connected by an edge and every edge connect a node of one group to a node of the other group. In our model, there is no link between the individuals or between the transactions but the only possible links are between an individual and a transaction. Note that this model is a simple re-writing of the corpus and does not make any historical assumption on the corpus itself. However, this model is restricted to transactions established before year 1500 to avoid an artefact due to the low ratio of digitized transactions posterior to 1500.

Understanding the reasons behind common contracting between two persons is one of the important historical issues related to such databases. Previous work on the corpus already emphasized a clear organization by dates that can be useful to automatically detect transcription errors. In this paper, our purpose is to measure how strong the influence of the spatial location is, in such a relational network. In order to understand the evolution of the impact of the spatial information through time, the original corpus is divided into three temporal parts:

— the first one lasts from 1250 to 1350: it approximatively corresponds to the period of the highest activity, before the Hundred Years' War and the Black Death. The network model derived from the transactions dated at that time contains 6174 nodes (2295 individuals and 3879 transactions) and, in the following, only the largest connected component is considered (2173 individuals and 3780 transactions);
— the second one lasts from 1350 to 1450 and approximately corresponds to the Hundred Years' War and a period of very low activity. The network model derived from the

corresponding transactions contains 1717 nodes and its largest connected component has 1441 nodes (509 individuals and 2362 transactions);
— the last one lasts from 1450 to 1500 and corresponds to the post Hundred Years' War period. At this moment, the number of transactions is increasing again. The network model derived from the corresponding transactions contains 2516 nodes (1099 individuals and 1417 transactions).

Using a force directed algorithm as described in (Fruchterman and Reingold, 1991), the three bipartite graphs are displayed in Figure 1. In this figure, the transactions are colored according to their parishes and the visual effect is that the network structure seems to be somehow correlated to the geographical labels, especially in the network corresponding to the first period (before the Hundred Years' War, top right of the figure). However, this visual assumption has to be confirmed. To that purpose, it is relevant to consider the information about the parish not only as a label but as a spatial information.

To do so, our proposal is to compute two types of "distances" between individuals:

— a "social distance" based on common citations in a same transaction. This "social distance" is computed from the definition of a projected graph (Section 3.1) by using three graph based similarities/dissimilarities;
— a "geographical distance" based on the "average spatial location" of each individual. This distance can cope with multiple spatial locations labeling.

These two "distances" are then compared by using a permutations based test in which the peculiar relations between the projected network and its spatial labeling is preserved.

3 Methodology

3.1 Projected graph

From the original relational model described in Section 2, a second (projected) graph can be derived to represent the "social" relations between the individuals in the network: the nodes of the projected graph are the individuals of the original graph and two nodes are linked by an edge if the two individuals have been actively involved in the same transaction. These edges can be weighted in a very intuitive manner, by the number of common transactions in which the individuals are common active participants. The Figure 2 illustrates the definition of such a projected network: in this figure, transactions are displayed with orange nodes and individuals with yellow numbered nodes.

To be able to measure the strength of the correlation between common contracting and close geographical locations, several similarity/dissimilarity indexes are defined in the following sections. In Section 3.2, similarities and dissimilarities are defined that measure the tendancy of two individuals to be involved in common transactions. Then, in Section 3.3, a dissimilarity measuring the distance between a typical location of each individuals is proposed. Finally, Section 3.4 describes a methodology to test if these two measures of proximity between pairs of individuals are significantly correlated or not.

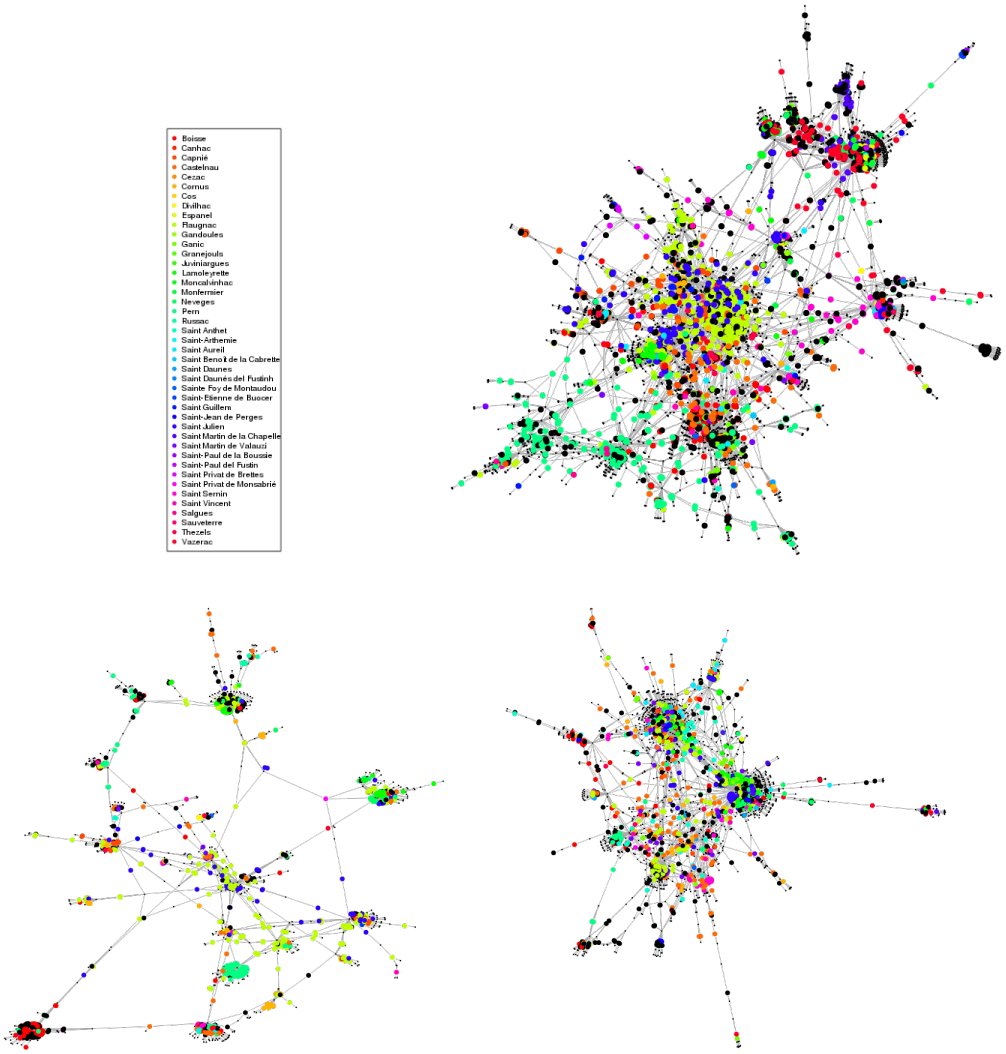


FIG. 1 – *Bipartite network transactions/individuals: small nodes are individuals and big ones are transactions. When it is known, the color of the transaction indicates its parish. Black is used when the information about the parish is not available. Top right: network derived from transactions dated between 1250 and 1350; bottom left: network derived from transactions dated between 1350 and 1450; bottom right: network derived from transactions dated between 1450 and 1500.*

3.2 Social "distances"

The dissimilarities described in this section are meant to quantify the proximity between two individuals viewed from the point of view of shared transactions: the higher they are, the

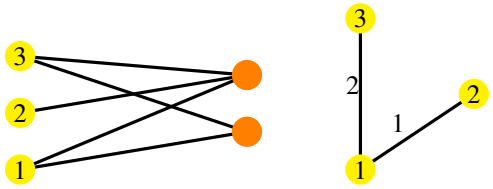


FIG. 2 – *Bipartite graph (left) and corresponding projected network for the yellow nodes*

"closer" the corresponding nodes are, according to the structure of the projected network defined in Section 3.1. In the following, these measures will be refereed as "social distances" (even though they are not, mathematically speaking, distances, but similarities or dissimilarities).

The simplest proposition is to use the adjacency matrix, $A = (A_{ij})_{i,j=1,\dots,N}$, of the unweighted projected graph defined in Section 2:

$$A_{ij} = \begin{cases} 1 & \text{if nodes } i \text{ and } j \text{ are actively involved in at least one common transaction} \\ 0 & \text{otherwise} \end{cases}$$

with, by convention, $A_{ii} = 0$. The choice of the unweighted version compared to the weighted version means that two persons are simply said to be "socially related", whatever the number of common transactions they share. This choice is driven by the need to leverage the fact that sometimes, a large number of very similar transactions (made in a same act) involved the same persons. For instance, a large exchange, between Arnaud Bernard Perarede and Raimond Perarede, resulted in a few acts with about thirty transactions that were not an indication that these two persons actually had a very strong relationship.

This similarity measure is easy to understand but can be somehow too crude to really capture a precise information on the proximity between two individuals in the network: two persons who have never been involved in the same transaction but are frequently connected to a same third person are treated similarly than two persons who have never been involved in the same transaction and have no common neighbors in the projected network. To soften this proximity index, two strategies are investigated:

— the length of the shortest path between two nodes in the projected network. This index is a dissimilarity measure of the social proximity between individuals, entirely based on the network structure;

— the use of a regularized version of the Laplacian matrix, that is strongly related to the graph structure (von Luxburg, 2007). The Laplacian of a graph is the matrix $L = (L_{ij})_{ij}$ where

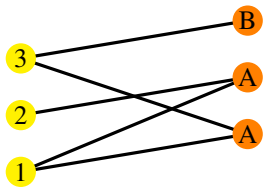
$$L_{ij} = \begin{cases} d_i & \text{if } i = j \\ -A_{ij} & \text{if } i \neq j \end{cases}$$

with $d_i = \sum_j A_{ij}$ and several kernels have been derived from this matrix, whose algebraic properties are strongly related to the graph structure (Smola and Kondor, 2003; Kondor and Lafferty, 2002). These kernels are natural similarity measures between nodes of the graph because it can be proved that they are dot products of an embedding of the graph in a (not explicitly defined) Hilbert space (Aronszajn, 1950). For its simplicity, the commute time kernel, introduced in (Fouss et al., 2007), is used as a social

similarity measure between two individuals. The commute time kernel computes the average time needed to reach a node from another one by a random walk on the graph. It is thus expected to have a meaning similar to that of the shortest path, even though it is probably smoother and less sensitive to false edges than the previous one. Many other similarity measures might have been defined from the adjacency matrix, simply noting that this matrix can be viewed as a matrix with individual observations of binary variables (Hubalek, 1982).

3.3 Geographical distances

Each individual is localized by the transactions in which he is actively involved. More precisely, the contingency table between individuals and parishes, T, can be computed as illustrated in Figure 3: in this figure, individuals are represented by orange nodes and transactions by yellow nodes. Transactions are labeled according to the location of the land concerned by the transaction.



	A	B
1	2	0
2	1	0
3	1	1

FIG. 3 – *Bipartite graph (left) and corresponding contingency table for the location labels (right)*

Then, the parishes' coordinates (latitude and longitude), C, are used to calculate an averaged location for each individual:

$$L = TC.$$

As some transactions are not spatially localized, some individuals end up with no location. To address this issue, a clustering based on the social network is performed using the modularity optimization as described in (Rossi and Villa-Vialaneix, 2011) [2]. Individuals with no location (about 12% to 17%, depending on the network) are then arbitrarily localized at the average location computed over the members of the same cluster. This imputation was performed either for the original spatial information and for the permuted locations.

Finally, a geographical distance between individuals is defined based on this knowledge: it is the euclidean distance between the individuals' averaged location coordinates.

3.4 A permutation test

This section describes the permutation test that has been used to test the correlation between the social distances and the geographical distances. This test is based on an idea similar to that of the Mantel test. The Mantel test is a statistical test of the correlation between two matrices. Indeed, because distances are not independent of each other, the relationship between two

2. Using the Java program available at `http://apiacoa.org/software/graph/index.en.html`

distance matrices can not be assessed by a standard test based on the correlation coefficient. The main idea of the Mantel test is to randomly permute the rows and columns of one of the matrices and to compare the observed correlation coefficient between the two matrices to the distribution of the correlation coefficients obtained from the permuted matrices.

In our context, such an approach is not accurate: being built from the same bipartite graph, the matrices that model the social and the geographical distances are correlated as a consequence. More precisely, if two individuals are involved in (almost) the same transactions, their geographical coordinates will be very close and their geographical distance will be very small. Hence, the more two individuals are linked by common transactions, the closer they are, both on a social point of view but also, on a geographical point of view.

Hence, the standard approach would always provide a significant correlation. The original approach is thus adapted to our particular framework. A permutation test is also used, to describe the null hypothesis in that context:

1. permute at random the spatial labels among the transactions in the bipartite network;
2. build (one of) the social distance matrix(ces) and the geographical distance matrix as described in Sections 3.1 to 3.3;
3. compute the correlation between the two matrices.

Hence, if two individuals share exactly the same transactions, their geographical distance is null whatever the permutation and this will not result in a change of the value of the correlation matrix. Hence, the null distribution of the correlation matrix is intended to assess the similarities between spatial information and social interactions beyond solely common transactions.

Two thousands such permutations were performed to estimate the distribution of the correlation coefficients, before, during and after the Hundred Years' War, under the null assumption

The locations are randomly distributed among the transactions.

The p-value of the correlation coefficient is thus derived from this analysis to understand the impact of the geographical location on the way the individuals interact together.

4 Results and discussion

The results of the permutation test described in Section 3.4 are given in Figures 4 to 6. For all the social distances described in Section 3.2, the value of the correlation between social and geographical distance is significantly different from that obtained with a random distribution of the geographical locations among the transactions: this result is not surprising in that context but the large difference between the observed value and the distribution obtained from random permutations is an indication of the strength of this correlation. The geographical location might have a strong influence on the way people interact, as described by the transactions in the corpus. Similar conclusions were already obtained from a different methodology in (Laurent and Villa-Vialaneix, 2011). Also, there is no noticeable differences between the three periods: the observed influence of the geographical location is always stronger than those observed on the two thousands correlation coefficients resulting from the permutations, before, during and after the Hundred Years' war. The gap between the observed value and the null distribution seems to be larger before the war than during it but the size of the two networks are not comparable so this might be simply a size effect.

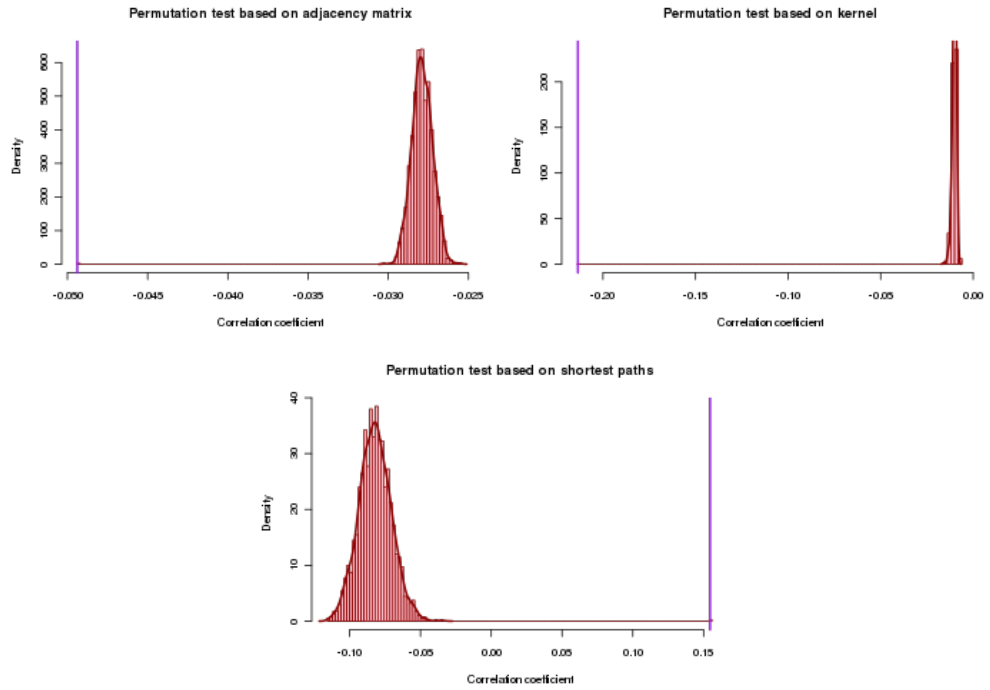


FIG. 4 – **Before the Hundred Years' War** - *Empirical distribution (pink histogram) and observed value (purple vertical line) of the test statistics to measure the correlation between the geographical distance and the social distance, based on the adjacency matrix (top left), the commute time kernel (top right) or the shortest path (bottom)*

The use of various ways to quantify the social distance leads to the same conclusion: the social distance has a correlation significantly larger with the geographical distance than in the random settings. But that conclusion has different scales, depending on which social distance is used. The correlation matrix and the commute time kernel are similarities and thus have a negative correlation to the geographical distance. The commute time kernel shows a larger difference between the observed statistics and the empirical distribution of the correlation coefficient than the adjacency matrix. Actually, the empirical distribution of the correlation between the commute time kernel and the geographical distance is almost centered around zero. On the contrary, the correlation between the length of the shortest path (that is a similarity) and the geographical distance is positive.

5 Conclusion

In this paper, we were able to illustrate a permutation based methodology to assess the significativity of a correlation between a network and spatial labels related to its nodes, when both the network and the labels are built from the same source. The proposed test was applied

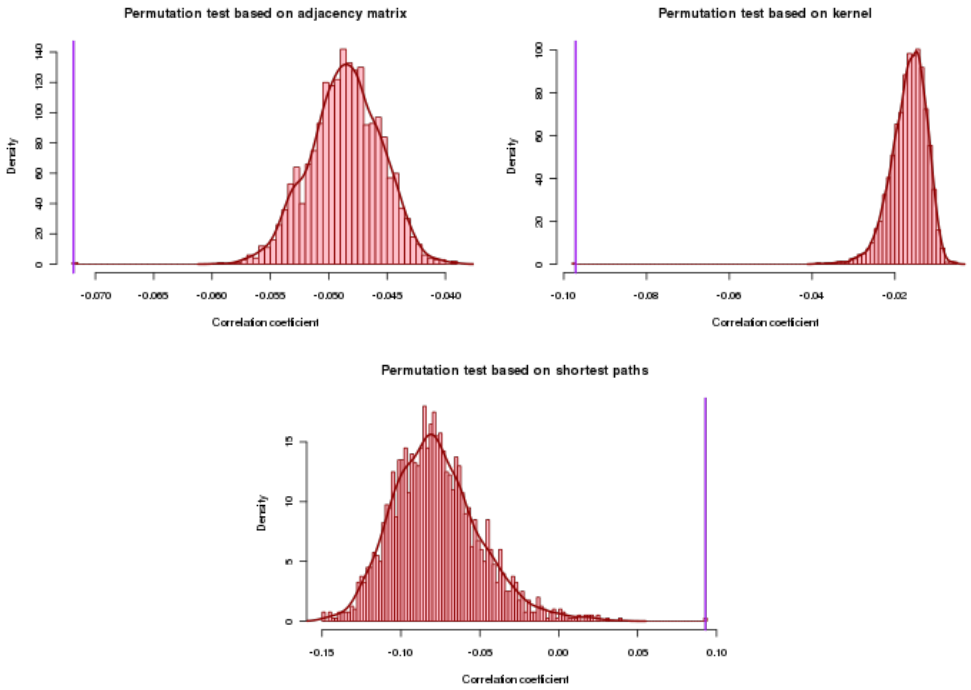


FIG. 5 – **During the Hundred Years' War** - *Empirical distribution (pink histogram) and observed value (purple vertical line) of the test statistics to measure the correlation between the geographical distance and the social distance, based on the adjacency matrix (top left), the commute time kernel (top right) or the shortest path (bottom)*

to a network built from a corpus of medieval notarial acts and proved the strong impact of the geo-localization on the way the persons interacts in the studied transactions. Different ways to quantify the similarities between nodes in the networks were used, all leading to the same conclusion. One way to go further in the study could be to modify the definition of the geographical distances, in order to take into account the lines of communication, terrain, economical proximities,...

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society 68*(3), 337–404.

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

Dorogovtsev, S. and J. Mendes (2003). *Evolution of Networks. From biological Nets to the Internet and WWW*. Oxford University Press.

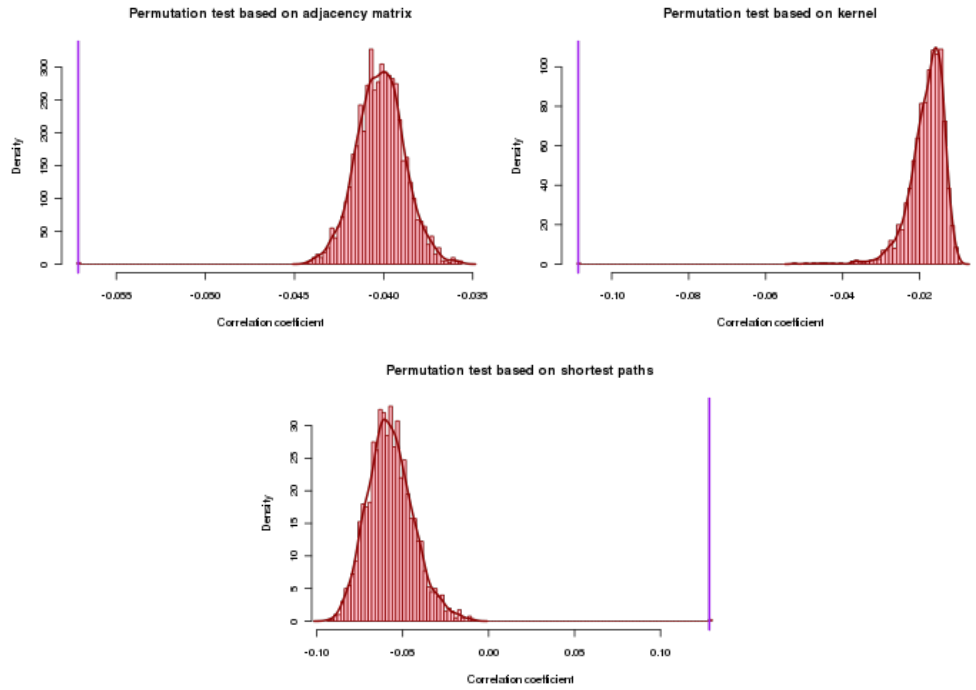


FIG. 6 – **After the Hundred Years' War** - *Empirical distribution (pink histogram) and observed value (purple vertical line) of the test statistics to measure the correlation between the geographical distance and the social distance, based on the adjacency matrix (top left), the commute time kernel (top right) or the shortest path (bottom)*

Fouss, F., A. Pirotte, J. Renders, and M. Saerens (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.

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

Hautefeuille, F. and B. Jouve (2012). Defining rural elites in the 13th to 15th centuries at the crossroads of historical, archeological and mathematical approaches. Submitted to the Journal of Interdisciplinary History.

Hubalek, Z. (1982). Coefficients of association and similarity based on (presence, absence): an evaluation. *Biological Review 57*, 669–689.

Kondor, R. and J. Lafferty (2002). Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 315–322.

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

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research 27*(2), 209–222.

Miquel, D. (2011). Analyse statistique d'un rÃl'seau social de la paysannerie médiévale. Technical report, UniversitÃl's Toulouse 1 & 2. Master's Thesis.

Moran, P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 10*, 243–251.

Newman, M., A. Barabási, and D. Watts (2006). *The Structure and Dynamics of Networks*. Princeton University Press.

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

Rossi, F., N. Villa-Vialaneix, and F. Hautefeuille (2012). Exploration of a large database of French notarial acts with social network methods. Submitted for publication.

Smola, A. and R. Kondor (2003). Kernels and regularization on graphs. In M. Warmuth and B. Schölkopf (Eds.), *Proceedings of the Conference on Learning Theory (COLT) and Kernel Workshop*, Lecture Notes in Computer Science, pp. 144–158.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing 17*(4), 395–416.

Résumé

Dans cet article, nous nous intÃl'ressons Ãă l'influence de la localisation spatiale des individus sur la maniÃÍre dont ils interagissent : le cas d'Ãl'tude que nous abordons est celui d'un corpus d'actes notariÃl's du Moyen-ĂĆge. Dans ce corpus, les individus interagissent par le biais de transactions communes qui sont gÃl'o-localisÃl'es. Ce travail cherche Ãă quantifier l'impact de la localisation gÃl'ographique sur le rÃl'seau de relations des individus. Comme positionnement gÃl'ographique et relations des individus sont construits Ãă partir de la mÃłme source (la liste des transactions auxquelles chaque individu a participÃl'), un test de Mantel classique (Mantel, 1967) est inadÃl'quat : nous proposons une mÃl'thodologie basÃl'e sur une adaptation du test de comparaison de distances que nous illustrons dans ce contexte.