Spatial correlation in bipartite networks

Nathalie Villa-Vialaneix
http://www.nathalievilla.org
nathalie.villa@univ-paris1.fr

MASHS 2012 - 4 Juin 2012

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

Collaboration

commencée dans le cadre du projet ANR Graph-Comp. [1]

1. http://graphcomp.univ-tlse2.fr

Plan

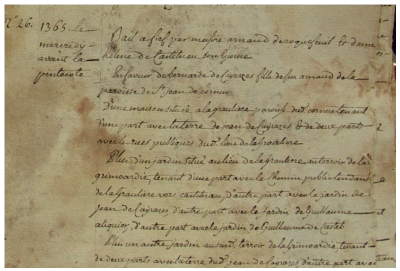
Présentation du corpus documentaire

Un grand corpus d'actes notariés médiévaux

Corpus provenant de l'œuvre d'un feudiste et conservé aux *archives du Lot (Cahors)*

- actes notariés relatifs à des rentes (principalement des baux à fief) ;
- établis entre 1250 et 1500 ;
- dans la seigneurie de Castelnau Montratier.

Description d'un acte type



Chaque acte est composé d'**une ou plusieurs transactions** (3 356 actes et 6 745 transactions saisies dans une base de données en libre accès à `http://graphcomp.univ-tlse2.fr`)

AD 46 48 J6 page 37, acte 26 1365, le mercredi avant la Pentecôte Bail à fief par messire Arnaud de Roquefeuil et Dame Hélène de Castelnau son épouse en faveur de Bernarde de Cayrazes, fille de feu Arnaud, de la paroisse de St Jean de Cornus, d'une maison située à La Graulière, paroisse du dit Cornus, tenant d'une part avec la terre de Jean de Cayrazes et de deux parts avec les rues publiques du dit lieu de La Graulière.

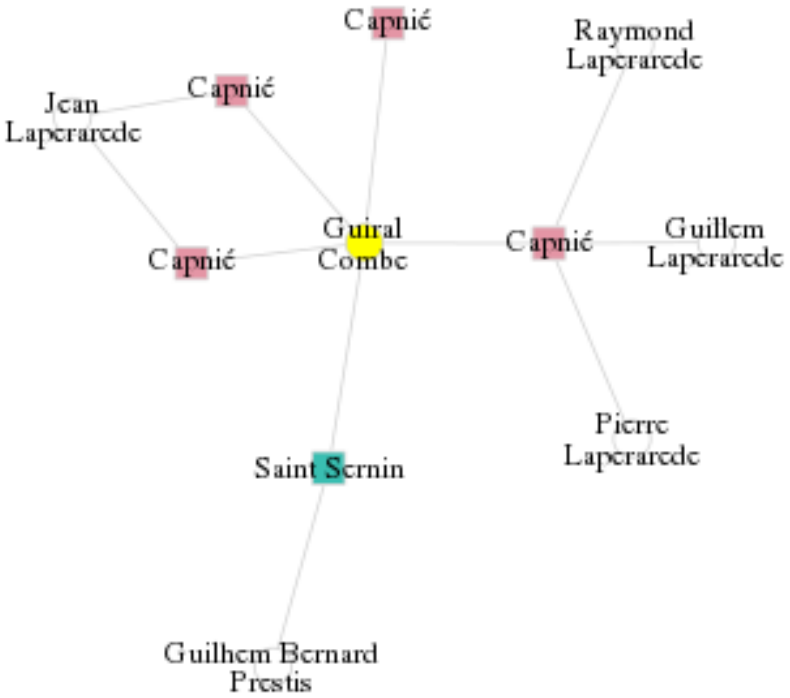
[...] 7 autres transactions...

sous la redevance de deux sous cahorcin d'acapte à mutation de seigneur ou de feudataire et de 3 emines d'avoine, l'emine vaut demi-setier et le setier 4 quartes et 1 poule à la notre Dame de septembre. Jean de Combelcau, notaire et commissaire d'autorité de monsieur l'official de Cahors.

Informations (+/- précises) sur : les personnes (tenanciers, seigneurs) impliqués, le notaire, les confronts, la date, le lieu, la rente...

Modélisation des données relationnelles

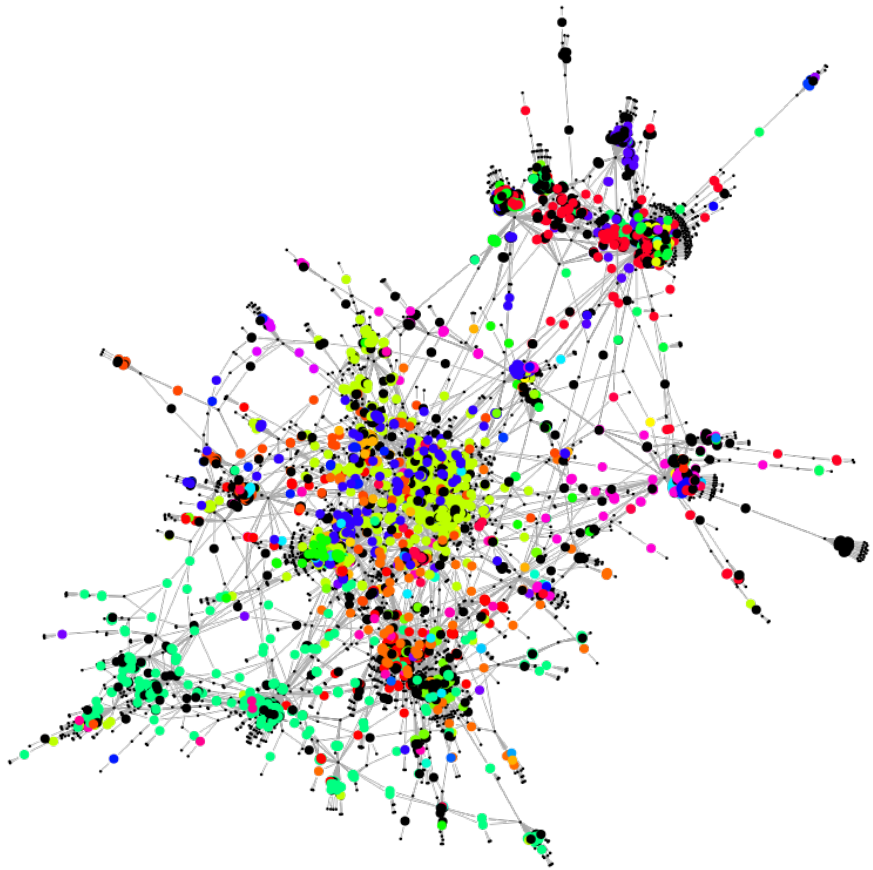
À partir des informations saisies dans la BD, définition d'un **modèle relationnel** (graphe *biparti*) :



- sommets : transactions et individus (3 918 sommets)
- arêtes : un individu est directement impliqué dans une transaction (6 455 arêtes)
- étiquettes (seulement transactions) : localisation géographique (paroisse)

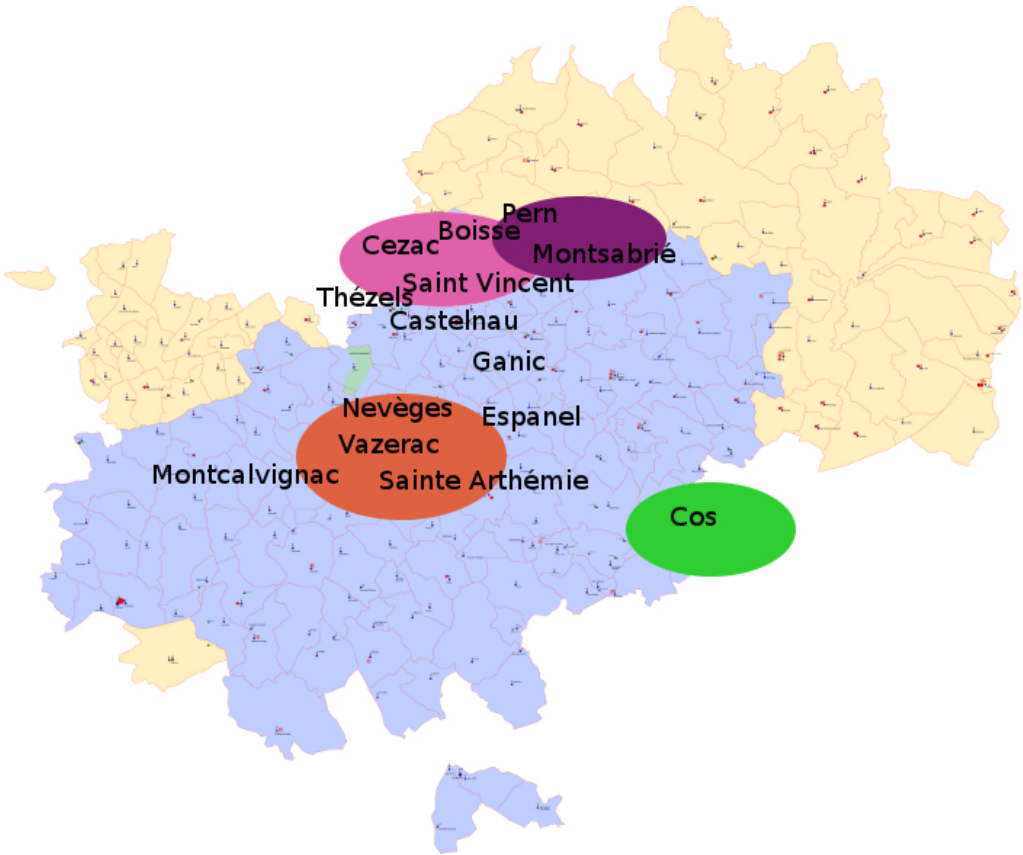
Données retenues

Étude de la **plus grande composante connexe** du modèle relationnel pour les transactions **entre 1250 et 1350**.



2 173 individus et 3 780 transactions (5 953 sommets)

Localisations géographiques des transactions



- 45 paroisses géo-localisées ;
- 985 transactions sans information de paroisse (environ 1/4 des transactions et 10% des individus pas du tout localisés) ;
- problème pour deux paroisses portant le même nom et exclues du traitement.

Problématiques

1. classification non supervisée de sommets pour repérer des groupes fortement connectés **[Boulet et al., 2008]** ;
2. outils de fouille de données pour l'**exploration** du graphe $\Rightarrow$ **repérer des erreurs de transcription [Rossi et al., 2012]** ;

Problématiques

1. classification non supervisée de sommets pour repérer des groupes fortement connectés **[Boulet et al., 2008]** ;
2. outils de fouille de données pour l'**exploration** du graphe ⇒ **repérer des erreurs de transcription [Rossi et al., 2012]** ;
3. **ICI** : comprendre l'**impact de la localisation géographique sur l'organisation des relations**.

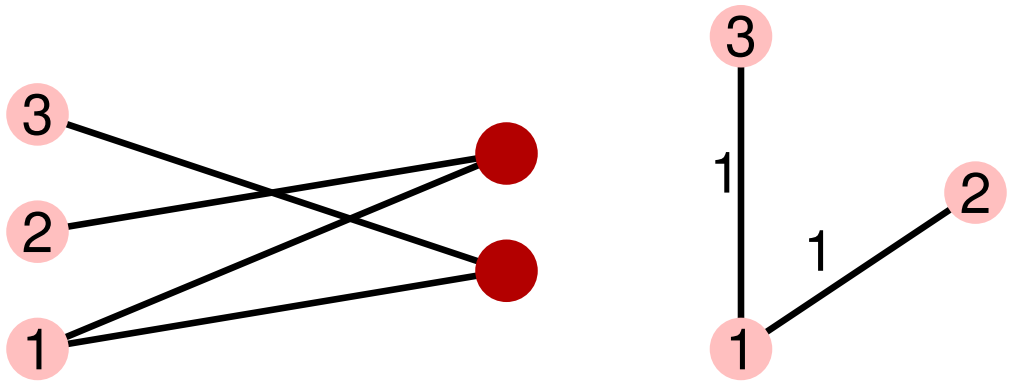
> *La manière dont les individus interagissent par le biais de transactions communes est-elle fortement liée aux localisations géographiques des biens concernés par les transactions ?*

Plan

Graphes projetés

À partir du graphe biparti, on peut définir :

- le **graphe des individus** : les sommets sont les 2 173 individus et les arêtes correspondent au fait que deux individus aient été directement impliqués dans au moins une transaction commune (pas de pondération) ;
- le **graphe des transactions** (pas utilisé) : les sommets sont les 3 780 transactions et les arêtes correspondent au fait que deux transactions ont été effectuées par le même individu.



Distances “sociales” entre les individus $(S_{ij})_{ij}$

Quantifier la proximité entre individus selon les transactions communes

Principe : Utiliser le graphe projeté des individus comme base de la mesure de leurs proximités (au sens “interagissent ensemble du point de vue des transactions”)

Distances “sociales” entre les individus $(S_{ij})_{ij}$

Quantifier la proximité entre individus selon les transactions communes

Principe : Utiliser le graphe projeté des individus comme base de la mesure de leurs proximités (au sens “interagissent ensemble du point de vue des transactions”)

Distances utilisées :

1. **longueur du plus court chemin** sur le graphe (dissimilarité) ;

Distances “sociales” entre les individus $(S_{ij})_{ij}$

Quantifier la proximité entre individus selon les transactions communes

Principe : Utiliser le graphe projeté des individus comme base de la mesure de leurs proximités (au sens “interagissent ensemble du point de vue des transactions”)

Distances utilisées :

1. **longueur du plus court chemin** sur le graphe (dissimilarité) ;
2. **similarités basées sur la matrice d’adjacence** :

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

Distances “sociales” entre les individus $(S_{ij})_{ij}$

Quantifier la proximité entre individus selon les transactions communes

Principe : Utiliser le graphe projeté des individus comme base de la mesure de leurs proximités (au sens “interagissent ensemble du point de vue des transactions”)

Distances utilisées :

1. **longueur du plus court chemin** sur le graphe (dissimilarité) ;
2. **similarités basées sur la matrice d’adjacence** : **version régularisée du Laplacien** ($L = \mathrm{Diag}(d_i)_i - A$ où d_i est le degré du sommet i) :

$$L^+$$

“commute time kernel” qui s’interprète comme le temps moyen pour joindre deux sommets par une marche aléatoire sur le graphe.

Distances “géographiques” entre individus

Quantifier la proximité géographique entre individus $(G_{ij})_{ij}$

Principe :

- Les individus sont localisés par la liste des paroisses des transactions dans lesquelles ils sont acteurs ;
- Les paroisses sont géo-localisées (latitude/longitude).

Distances “géographiques” entre individus

Quantifier la proximité géographique entre individus $(G_{ij})_{ij}$

Principe :

- Les individus sont localisés par la liste des paroisses des transactions dans lesquelles ils sont acteurs ;
- Les paroisses sont géo-localisées (latitude/longitude).

Positionnement “type” : on géo-localise un indivu (latitude/longitude) à sa position moyenne sur l’ensemble des paroisses.

Distances “géographiques” entre individus

Quantifier la proximité géographique entre individus $(G_{ij})_{ij}$

Principe :

- Les individus sont localisés par la liste des paroisses des transactions dans lesquelles ils sont acteurs ;
- Les paroisses sont géo-localisées (latitude/longitude).

Positionnement “type” : on géo-localise un indivu (latitude/longitude) à sa position moyenne sur l’ensemble des paroisses.
Distance “géographique” : la distance géographique entre deux individus est alors la distance euclidienne entre leurs positionnements types respectifs.

Corrélation entre matrices de distances

Question : Y a-t-il une corrélation significative entre la distance “sociale” et la distance “géographique” ?

> *Est-ce que le positionnement géographique des individus, vu au travers de leurs transactions, a une influence significative sur le choix des personnes avec lesquelles ils interagissent ?*

Corrélation entre matrices de distances

Question : Y a-t-il une corrélation significative entre la distance “sociale” et la distance “géographique” ?

> *Est-ce que le positionnement géographique des individus, vu au travers de leurs transactions, a une influence significative sur le choix des personnes avec lesquelles ils interagissent ?*

Rappels sur les tests de corrélations

Calcul du **coefficient de corrélation** entre $(S_{ij})_{ij}$ et $(G_{ij})_{ij}$, Cor : comment tester

> *Cor n'est pas significativement plus élevée (en valeur absolue) que si les individus interagissent indépendamment de leurs positionnements géographiques*

?

Corrélation entre matrices de distances

Question : Y a-t-il une corrélation significative entre la distance “sociale” et la distance “géographique” ?

> *Est-ce que le positionnement géographique des individus, vu au travers de leurs transactions, a une influence significative sur le choix des personnes avec lesquelles ils interagissent ?*

Rappels sur les tests de corrélations

Calcul du **coefficient de corrélation** entre $(S_{ij})_{ij}$ et $(G_{ij})_{ij}$, Cor

Problème : Les distances ne sont pas indépendantes (si i est proche de j et j proche de k alors i est proche de k) donc le test de corrélation habituel n’est pas valable.

Corrélation entre matrices de distances

Question : Y a-t-il une corrélation significative entre la distance “sociale” et la distance “géographique” ?

> *Est-ce que le positionnement géographique des individus, vu au travers de leurs transactions, a une influence significative sur le choix des personnes avec lesquelles ils interagissent ?*

Rappels sur les tests de corrélations

Calcul du **coefficient de corrélation** entre $(S_{ij})_{ij}$ et $(G_{ij})_{ij}$, Cor

Approche de Mantel : tester la significativité de la corrélation entre deux matrices de distances... Répéter P fois (P grand) :

- permuter D fois, aléatoirement des lignes et des colonnes de S ;
- calculer le coefficient de corrélation sur la base de la matrice dont les lignes/colonnes ont été permutées : Cor^p.

puis comparer la valeur Cor observée à la distribution des valeurs sous l’hypothèse nulle $(\mathrm{Cor}^p)_{p=1,\dots,P}$.

Adaptation du test de Mantel

Pourquoi le test de Mantel est inadapté ici ? $(S_{ij})_{ij}$ et $(G_{ij})_{ij}$ sont construits à partir du **même graphe biparti** $\Rightarrow$ elles sont naturellement corrélées.

Adaptation du test de Mantel

Pourquoi le test de Mantel est inadapté ici ? $(S_{ij})_{ij}$ et $(G_{ij})_{ij}$ sont construits à partir du **même graphe biparti** $\Rightarrow$ elles sont naturellement corrélées.

Test de permutations basé sur le graphe biparti

Répéter P fois (P grand)

1. permuter aléatoirement les étiquettes géographiques des transactions ;
2. déterminer la nouvelle distribution des paroisses des individus et leurs nouveaux positionnements “type”. En déduire la matrice des distances géographiques de la permutation aléatoire $(G^p_{ij})_{ij}$;
3. calculer le coefficient de corrélation Cor^p entre S et G^p.

Interprétation de la procédure

Interprétation

La distribution des $(\mathrm{Cor}^p)_{p=1,\ldots,P}$ correspond à la distribution des coefficients de corrélation sous l'hypothèse nulle

> *Les positionnements géographiques des lieux concernés par les transactions sont indépendants du choix des individus qui agissent dans ses transactions.*

où, par "indépendant", on entend "au sens géographique du terme", c'est à dire, au sens de la distance entre ces lieux géographiques (et pas seulement au sens du "nom" de la paroisse).

Interprétation de la procédure

Interprétation

La distribution des $(\mathrm{Cor}^p)_{p=1,\dots,P}$ correspond à la distribution des coefficients de corrélation sous l'hypothèse nulle

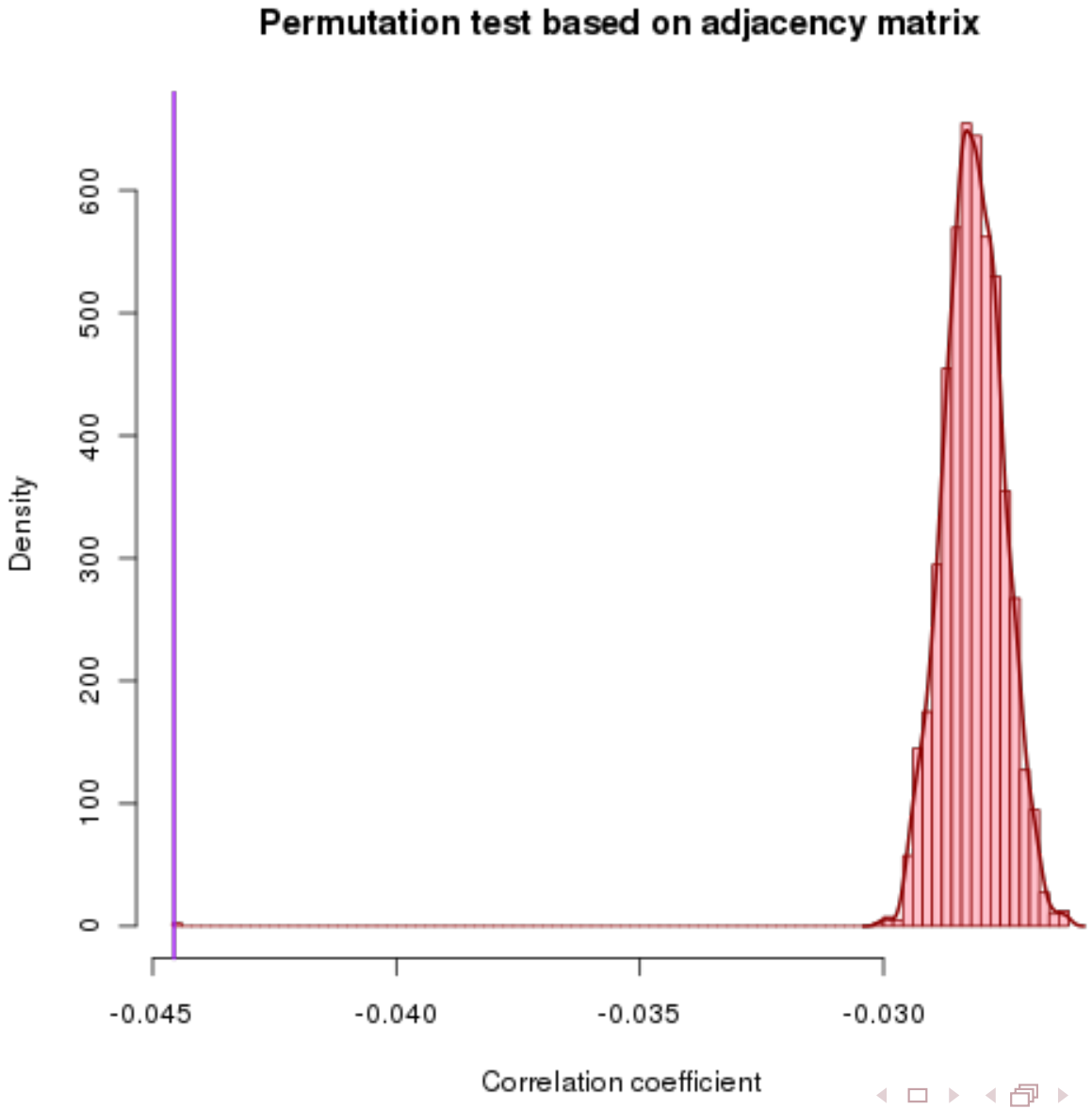
> *Les positionnements géographiques des lieux concernés par les transactions sont indépendants du choix des individus qui agissent dans ses transactions.*

où, par "indépendant", on entend "au sens géographique du terme", c'est à dire, au sens de la distance entre ces lieux géographiques (et pas seulement au sens du "nom" de la paroisse).

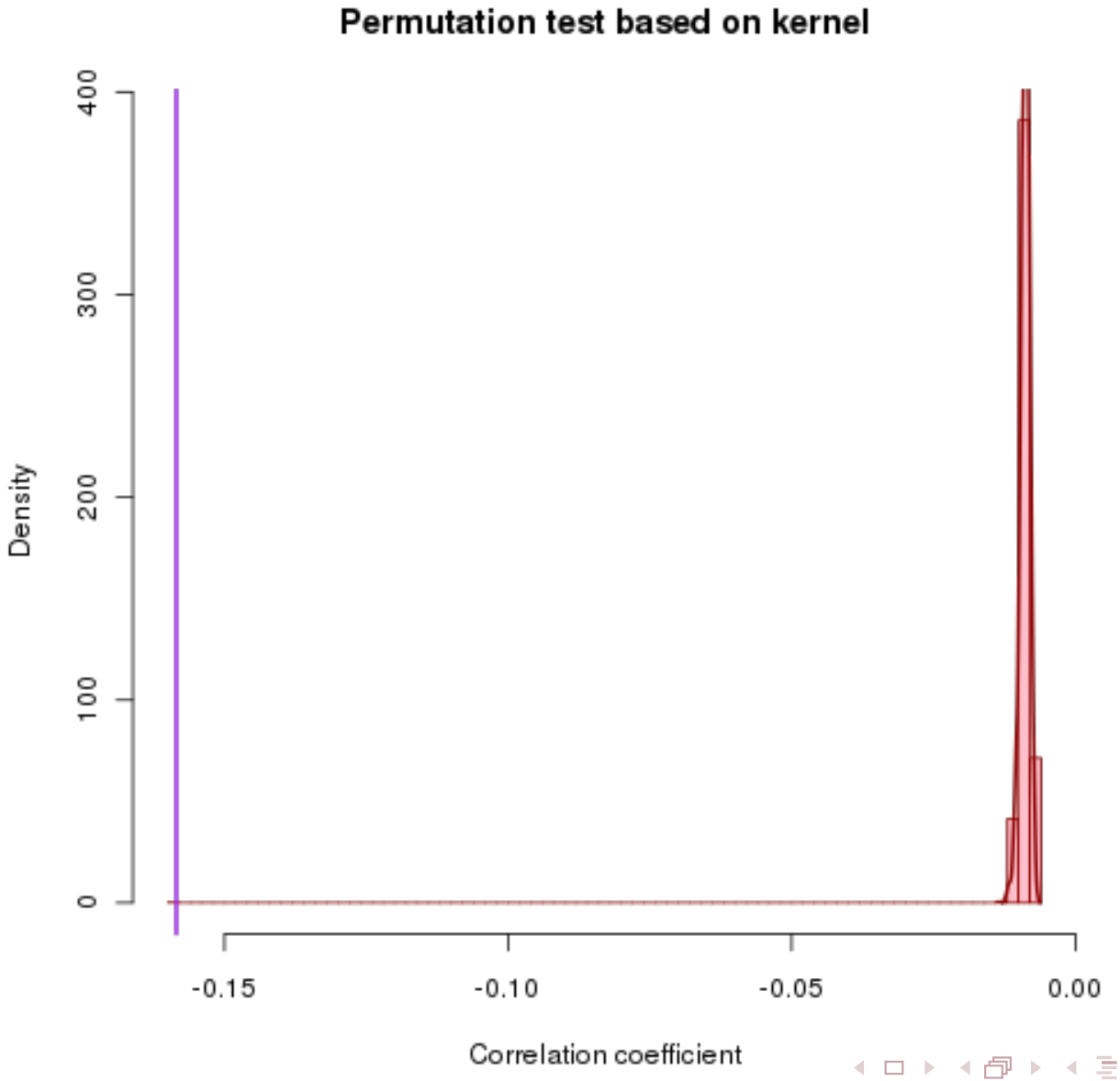
Conclusion : Si Cor est exceptionnellement élevé (ou faible) en comparaison de la distribution des $(\mathrm{Cor}^p)_p$, la localisation géographique a un impact significatif sur le choix des individus impliqués dans les transactions.

Plan

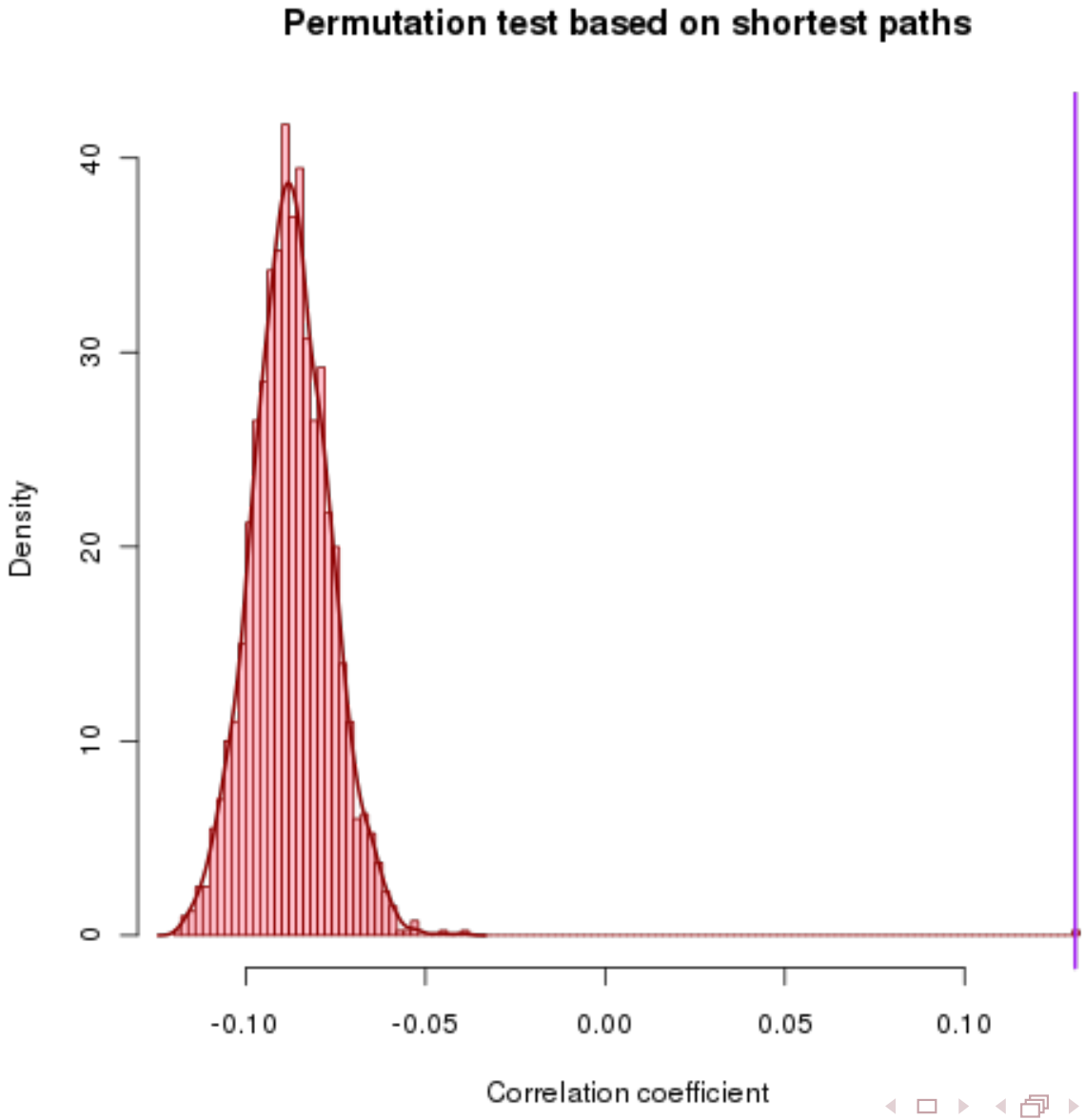
Résultats selon les diverses distances sociales choisies



Résultats selon les diverses distances sociales choisies



Résultats selon les diverses distances sociales choisies



Conclusions et perspectives

Conclusions

- résultats concordants pour les trois “distances” sociales : les individus qui interagissent dans les mêmes transactions ont des **positionnements géographiques très proches** ;
- le phénomène est encore plus marqué par l’utilisation d’une distance sociale “souple” (individus proches ≡ individus ayant des partenaires de transactions communs).

Conclusions et perspectives

Conclusions

- résultats concordants pour les trois “distances” sociales : les individus qui interagissent dans les mêmes transactions ont des **positionnements géographiques très proches** ;
- le phénomène est encore plus marqué par l’utilisation d’une distance sociale “souple” (individus proches ≡ individus ayant des partenaires de transactions communs).

Limites et perspectives

- **distances géographiques plus pertinentes ?** (utiliser la topologie du terrain, les routes ?)
- vérifier l’**effet de transactions multiples sur la même paroisse dans le même acte** (collecter la distribution des localisations à partir des actes plutôt que des transactions ?)
- utiliser cette méthodologie pour **étudier l’évolution de l’impact de l’implantation géographique sur les relations entre individus**.

Merci pour votre attention... Des questions ?

Boulet, R., Jouve, B., Rossi, F., and Villa, N. (2008).

Batch kernel SOM and related laplacian methods for social network analysis.
Neurocomputing, 71(7-9):1257–1273.

Rossi, F., Villa-Vialaneix, N., and Hautefeuille, F. (2012).

Exploration of a large database of French notarial acts with social network methods.
Submitted.