Exploration relationnelle d'un corpus d'actes notariés médiévaux[1]

Fabrice Rossi, Nathalie Villa-Vialaneix, Florent Hautefeuille

Introduction

Dans cet article, nous nous proposons de montrer comment construire, au moyen d'outils statistiques et mathématiques, différents points de vue sur une grande base de documents du Moyen-Âge, dans le but de faciliter l'extraction d'informations sur ses principales caractéristiques et de fournir des méthodes semi-automatiques de recherche d'erreurs de transcription éventuelles. Les documents constituant le corpus, des actes notariés, sont décrits par une base de données relationnelle qui inclut non seulement leurs caractéristiques propres (dates, lieux, biens concernés, etc.), mais également les descriptions des acteurs des actes (nom, statut social, métier, etc.), ainsi que les interactions entre ces deux types d'entités : cette approche donne une représentation du corpus non plus seulement comme un ensemble de documents et de (noms de) personnes mais aussi comme un modèle mathématique d'interaction, un graphe (aussi appelé communément un réseau) reliant les documents aux personnes qu'ils mentionnent et vice versa. Ce réseau reliant des entités de natures différentes, les documents et les personnes, donne naturellement naissance à deux autres réseaux : celui des documents, où la relation entre deux documents correspond à la présence d'une même personne dans ces deux documents, et celui des individus, où deux personnes sont dites « en relation » si elles sont impliquées dans une même transaction. Les trois réseaux ainsi construits donnent des visions alternatives et complémentaires (et simplifiées dans le cas de deux réseaux secondaires) du corpus, ce qui facilite l'extraction (semi)-automatique de faits importants ou atypiques.

Le graphe n'est pas utilisé ici comme support de représentation d'un « réseau social », ce qui demanderait de faire des hypothèses fortes (et discutables) sur la nature du lien social induit par le fait de dépendre du même seigneur ou d'avoir le même notaire, mais d'une manière à la fois plus générale et moins complexe, pour modéliser un ensemble de relations de nature informationnelle entre objets : ici le modèle principal est celui de la citation des noms de personnes dans des documents. De façon plus générale, cette stratégie s'applique à toute base de données relationnelle qui contient, comme son nom l'indique, un ensemble de relations de différentes natures entre des objets de différentes natures. Chaque type de relation peut être vu comme un graphe d'interaction entre les objets concernés : on peut songer, par exemple[2], à des bases de données de citations entre articles ou bien des bases de données de publications (la relation naturelle étant ici celle qui lie un auteur à un article).

Nous montrons comment, en combinant un modèle relationnel global sur le corpus documentaire avec des informations additionnelles décrivant à la fois les documents (date, lieu...) et les individus mentionnés dans ceux-ci (noms, âge...), il est possible de mobiliser des outils mathématiques développés dans le cadre de l'étude des réseaux sociaux pour extraire de l'information de ce corpus (et donc plus généralement de bases de données relationnelles). En particulier, une telle modélisation peut être utilisée pour extraire des erreurs de transcription des transactions sous forme numérique. Par ailleurs, lorsque le corpus est d'une taille telle qu'une exploration exhaustive

1 Ce travail a été financé en partie par l'ANR au travers du projet « Graphes-Comp », réf. ANR-05-BLAN-0229.

2 Des exemples de graphes comme outils pour modéliser des relations dans des applications courantes sont décrits dans Dorogovtsev Sergei N. and Mendes Jose F.F., *Evolution of Networks*, Oxford, Oxford University Press, 2003, p. 31-83.

manuelle de son contenu est difficile, les outils de fouille de données relationnelles permettent de résumer les principales caractéristiques du corpus : d'un point de vue macroscopique, la structure des relations peut être simplifiée par des méthodes de classification automatique, comme expliqué dans la suite de l'article. Changeant de point de vue, un zoom sur des phénomènes locaux particuliers, mis en exergue par des caractéristiques individuelles sur les sommets du graphe, permet de repérer des entités atypiques qui peuvent correspondre à des phénomènes intéressants ou bien à des erreurs dans la base de données. Enfin, les méthodes de visualisation dédiées à la représentation de graphes peuvent être utilisées pour proposer à l'utilisateur non mathématicien une configuration parlante et intuitive de son corpus : des informations de natures multiples peuvent être ajoutées à la représentation permettant de corréler de manière simple les relations entre entités avec les attributs décrivant ces entités.

L'approche est ici illustrée sur un grand corpus d'actes notariés médiévaux, contenant plusieurs milliers de transactions qui impliquent, au total, plusieurs milliers de personnes. Le corpus entier a été saisi dans une base de données relationnelle et contient des transactions de nature similaire qui ont toutes été effectuées dans une même petite seigneurie : le corpus est donc caractérisé par une homogénéité importante. De ce corpus a été extrait un graphe dont les sommets (plus de dix mille) sont les transactions et les personnes activement impliquées dans celles-ci et qui met en relation les personnes avec les transactions auxquelles elles ont participé. Le reste de l'article est organisé comme suit : la section « Données et modèle » décrit le corpus de manière plus précise, la base de données qui a été définie à partir de celui-ci et le modèle relationnel qui en a été extrait. La section « Analyse macroscopique » se penche sur l'analyse globale du réseau, à l'échelle macroscopique, en utilisant des indicateurs statistiques et des outils de visualisation et de classification non supervisée. La section « Analyse locale » change de perspective et montre comment l'utilisation d'indicateurs numériques locaux permet de mettre en valeur des personnes clés dans le graphe. Enfin, la section « Propager des informations » montre comment de la propagation d'information, combinant les différentes images du réseau en une configuration multiple, permet de proposer des outils de vérification semi-automatiques de la validité des saisies effectuées dans la base de données. La Conclusion résume les bénéfices de l'utilisation d'un modèle relationnel pour obtenir des représentations alternatives et complémentaires d'un corpus documentaire.

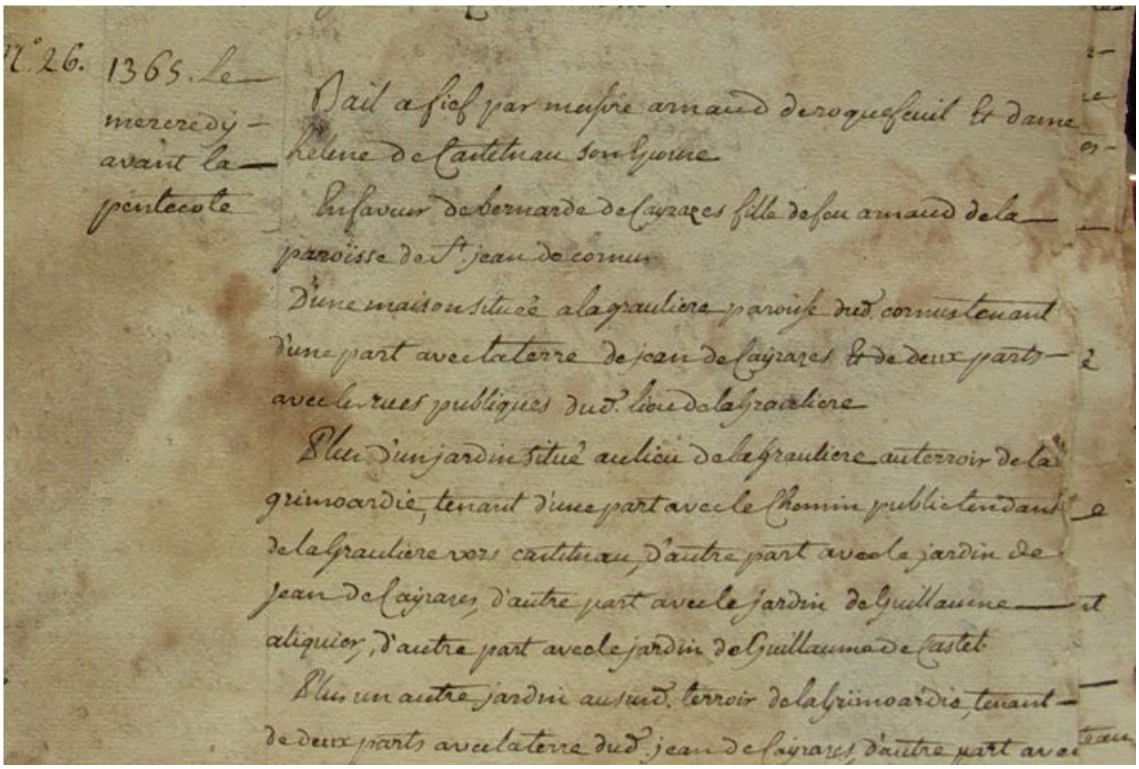
Données et modèle

Le corpus que nous étudions dans cet article provient des archives départementales du Lot (France) où il est consultable[3]. Il est divisé en quatre registres dont les cotes sont, respectivement AD 46 48 J 3, AD 46 48 J 4, AD 46 48 J 5 et AD 46 48 J 6. Il nous est parvenu grâce au travail d'un feudiste qui s'est employé, durant une vingtaine d'années au XVIII^{e} siècle, à collecter tous les actes notariés concernant les terres des seigneurs successifs de la seigneurie de Castelnau Montratier (Lot, France). Ce travail était mandé par le nouveau propriétaire de la seigneurie, Léon de Bonal, un bourgeois récemment anobli, pour lui permettre de réclamer des droits de rente sur ses terres nouvellement acquises. Le corpus est donc constitué d'un nombre important de documents dont les actes originaux ont été perdus mais qui ont pu nous parvenir grâce à ce travail de retranscription.

Les documents du corpus sont tous des actes notariés, chacun décrivant une ou plusieurs transactions et présentant un certain nombre de caractéristiques communes : tout d'abord, les transactions concernent des lieux situés sur la seigneurie de Castelnau Montratier, située près de l'actuel village du même nom (Lot, France). La seigneurie de Castelnau Montratier était constituée d'une quarantaine de villages, pour une superficie totale d'environ 300 km², ce qui en fait une zone géographique suffisamment réduite pour être considérée comme une unité homogène. Par ailleurs, toutes les transactions relevées par le feudiste décrivent des accords qui, bien que de natures

3 *Archives départementales du Lot*, ed. by Gérard Miquel and Willy Luis http://www.lot.fr/cg_archives.php.

différentes (vente, location, donation, bail à fief...), portent pour la plupart sur des terres et impliquent des rentes. Ces transactions ont été réalisées entre 1238 et 1768, avec une densité de transaction assez variable tout au long de la période. Ainsi, le corpus entier peut être considéré comme un ensemble très représentatif, si ce n'est exhaustif, des contrats portant sur des terres, qui ont été rédigés dans la seigneurie durant cette période.


N°26. 1365. Le mercredy avant la pentecote

Bail a fief par messire arnaud de roquefeuil et dame helene de castelnau son epouse
en faveur de bernarde de cayrazes fille de feu arnaud de la
paroisse de St. jean de cornus
d'une maison située a la graulière paroisse dud. cornus tenant
d'une part avec la terre de jean de cayrazes et de deux parts
avec les rues publiques dud. lieu de la graulière
Plus d'un jardin situé au lieu de la graulière au terroir de la
grimoardie, tenant d'une part avec le chemin public tendant
de la graulière vers castelnau, d'autre part avec le jardin de
jean de cayrazes, d'autre part avec le jardin de guillaume et
aliquior, d'autre part avec le jardin de guillaume de castet
Plus un autre jardin au dud. terroir de la grimoardie, tenant
de deux parts avec la terre dud. jean de cayrazes, d'autre part avec

Figure 1: Exemple d'acte (photographie partielle) contenant plusieurs transactions ; la première est un bail à fief, enregistré sous l'identifiant 142 dans la base de données. Figure reproduite avec l'aimable autorisation des Archives départementales du Lot, copyright Florent Hautefeuille, 2005.

Les transactions ont été saisies dans une base de données librement consultable en ligne[4]. Plus de 75 % des transactions du corpus entier ont ainsi été numérisées, avec un ratio différent selon le registre (la priorité a, en effet, été donnée à un sous-espace géographique précis et homogène de la seigneurie et aux transactions antérieures à 1500). En guise d'exemple, l'acte dont la photographie partielle est donné dans la Figure 1, contient la transaction suivante :

« 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 braulière, 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.

[…][5]

sous la redevance de deux sous cahorcin d'acapte à mutation de seigneur ou de feudataire et de 3

4 *Graph-Comp : Études des réseaux sociaux de sociabilités paysans au Moyen-Age dans la châtellenie Castelnau-Montratier*, ed. Florent Hautefeuille *et al.* http://graphcomp.univ-tlse2.fr.

5 7 autres transactions concernant deux jardins, un pré et quatre parcelles de terre.

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. »

Cette transaction est typique des transactions présentes dans la base de données : elle contient des informations multiples dont la référence de l'acte, AD 46 48 J6 page 37, acte 26 (qui est écrite dans la marge), la date de la transaction, le mercredi avant la Pentecôte, l'année 1365 (aussi écrite dans la marge), les (noms des) seigneurs directement concernés par la transaction qui sont *Arnaud de Roquefeuil* et *Dame Hélène de Castelnau* (son épouse), le nom du tenancier directement impliqué dans la transaction, *Bernarde Cayrazes*, la localisation de la terre objet de la transaction, le village de *La Graulière, paroisse de Cornus*, le nom du confront de cette terre, *Jean Cayrazes,* ainsi que le nom du notaire qui a rédigé l'acte, *Jean de Combelcau*. Une première configuration informationnelle du corpus peut être proposée au travers du modèle de base de données relationnelle utilisé pour enregistrer ces informations et dont une version très simplifiée est illustrée dans la Figure 2.

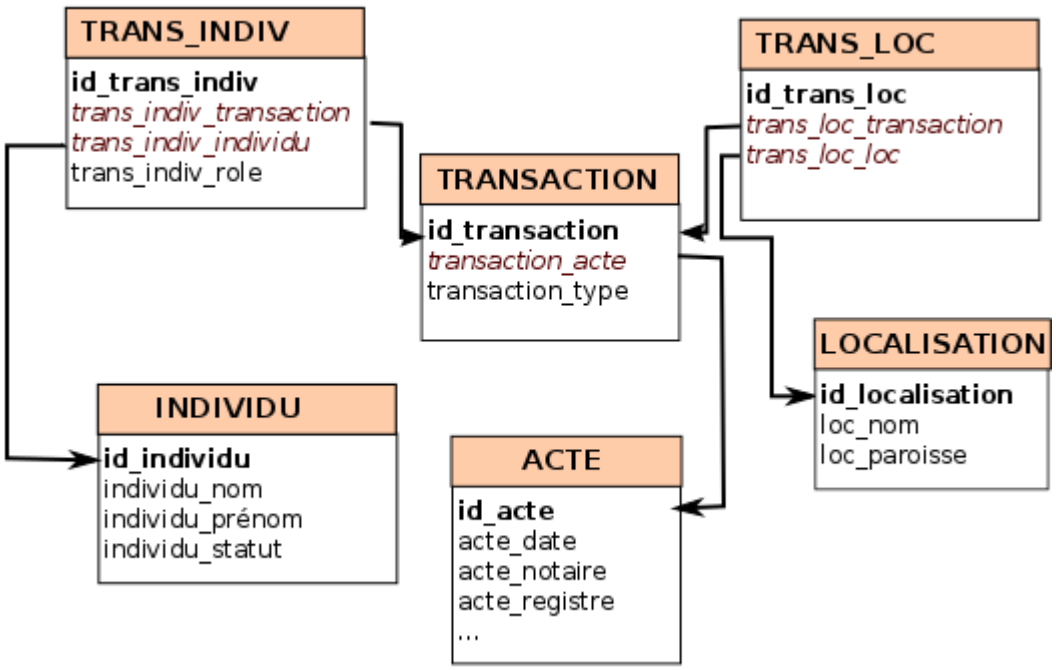


Figure 2: Modèle relationnel très simplifié de la base de données dans laquelle les transactions ont été numérisées.

Pour l'exemple de la transaction proposée dans la Figure 1, la transaction de l'acte *26* (*id_acte*), établi le *mercredi avant la Pentecôte 1365* (*acte_date*), par le notaire *Jean de Combelcau* (*acte_notaire*), provenant du registre *AD4648J6* (*acte_registre*) est un *bail à fief* (*transaction_type*) impliquant *Arnaud de Roquefeuil, Dame Hélène de Castelnau, Bernarde de Cayrazes* et *Jean Cayrazes* (*individu_prénom et individu_nom*), respectivement *seigneurs* (pour les deux premiers), *tenancier* (pour la troisième) et *tenancier confront* (pour le dernier) (*trans_indiv_role*). La transaction concerne une maison située à *La Graulière* (*id_localisation*), paroisse de *Cornus* (*loc_paroisse*). D'autres informations sur les qualificatifs relatifs aux individus, sur les rentes consenties, sur la nature de l'objet de la transaction, etc. ont été également numérisées mais nous ne les mentionnons pas ici afin de simplifier la présentation. Notons que nous utiliserons régulièrement le raccourci « individu » pour désigner en fait un nom de personne. Comme nous le verrons dans la suite de l'article, ce raccourci est commode mais masque le difficile problème des homonymies : il

ne faut pas y voir une volonté de minimiser ce problème qui est au contraire au cœur des analyses que nous proposons.

De cette base de données a été tirée une deuxième présentation du corpus, sous la forme d'un graphe. Dans ce réseau, les sommets modélisent les entités, transactions et individus directement impliqués dans une transaction. Les confronts ont été exclus du modèle car leur rôle est indirect et inactif, ce qui veut dire que la nature de relation entre un confront et une transaction n'est pas la même que celle de la relation entre un tenancier et une transaction : il faudrait de ce fait considérer un autre graphe. Les relations entre entités, ou arêtes, connectent des paires de sommets individu/transaction lorsqu'un individu est directement impliqué dans une transaction. Ce type de graphe est appelé *biparti*[6] car il ne connecte jamais deux sommets de même type (deux individus ou deux transactions). Une représentation d'une petite partie de ce modèle est donnée dans la Figure 7. Sur ces graphiques, un individu, appelé *Guiral Combe*, est impliqué dans cinq transactions (les dates des transactions ont été ajoutées sur le graphique de gauche et les paroisses des transactions sur le graphique de droite). Deux de ces transactions ont été réalisées avec une personne nommée *Jean Laperarede*, une autre avec *Guilhem Bernard Prestis*, une autre avec trois personnes, nommées *Pierre, Guillem* et *Raymond Laperarede* et la dernière est une transaction dans laquelle *Guiral Combe* est le seul participant actif.

Signalons, pour finir, que nous avons restreint notre modèle aux transactions effectuées avant 1500 pour ne pas biaiser la vision du corpus par le fait que seule une faible proportion de transactions postérieures à cette date ont été numérisées jusqu'à maintenant. Le graphe ainsi obtenu contient 10 542 sommets dont 6 487 transactions impliquant au moins un individu et 4 055 individus impliqués au moins dans une transaction comme tenancier ou comme seigneur actif (et pas seulement comme confront).

Analyse macroscopique

La modélisation du corpus par le biais d'un graphe peut être appréhendée selon divers points de vue, global, local ou bien encore en combinant plusieurs d'entre eux. Nous commençons ici par montrer comment une analyse macroscopique du réseau, effectuée par le biais du calcul d'indices et l'étude de caractéristiques propres aux réseaux, permet d'en comprendre des traits importants. Une approche courante consiste à commencer par l'analyse de la connectivité du graphe qui consiste à essayer de savoir si tout couple de sommets du graphe peut être relié par une suite d'arêtes (ce que l'on appelle un *chemin*[7] en théorie des graphes). Un tel graphe est dit *connexe* et tout graphe non connexe est constitué de plusieurs *composantes connexes*[8] qui sont des sous-graphes maximaux connexes ; les composantes connexes étant déconnectées les unes des autres, elles sont analysées séparément. Dans le cas de l'exemple étudié dans cet article, le graphe n'est pas connexe mais contient une composante connexe de grande taille (ce que l'on appelle souvent une *composante géante*) qui regroupe 3 755 individus et 6 270 transactions, soit 95,1 % des sommets du graphe initial. Pourtant, si on compare la couverture de la plus grande composante connexe à celle à laquelle on pourrait s'attendre dans un modèle de graphe aléatoire de même nature (graphe biparti avec une même distribution de degrés), la plus grande composante connexe du graphe étudié peut

6 Le vocabulaire relatif à la théorie des graphes peut être trouvé dans Voloshin Vitaly I., *Introduction to Graph Theory*, New York, Nova Science, 2009, ou bien, en termes moins formels, dans Scott John P., *Social Network Analysis: A Handbook*, London, Sage, « 2nd edition », 2000.

7 De manière plus précise, un chemin est une suite de sommets tel que chaque sommet de la suite partage une arête avec le sommet suivant dans la suite.

8 Les composantes connexes peuvent être obtenues de la manière suivante : partant d'une personne choisie au hasard, on collecte les transactions dans laquelle cette personne est impliquée puis les personnes également impliquées dans ces transactions et ainsi de suite jusqu'à avoir parcouru tous les sommets du graphe ou bien montré qu'une telle opération était impossible.

être considérée comme plutôt petite : des simulations effectuées par ordinateur[9] montrent, en effet, qu'en moyenne, la plus grande composante connexe, dans un modèle aléatoire, occupe 98,4 % des sommets du graphe initial. La conclusion de cette comparaison est que le graphe défini à partir du corpus d'actes notariés a une connectivité significativement plus faible que celle d'un graphe sans sous-structure : c'est le signe qu'il contient au contraire des sous-structures, qui se manifestent notamment par l'existence de zones de densités plus fortes et d'autres de densités plus faibles que dans un graphe présentant une répartition homogène des arêtes sur l'ensemble des paires de sommets. Les autres composantes connexes du graphe, au nombre de 107, sont de toutes petites tailles (moins de 11 personnes chacune) et ne seront donc pas étudiées dans la suite de cet article, puisqu'elles apportent très peu d'information chacune.

9 KANNAN Ravi, TETALI Prasad, VEMPALA Santosh, « Simple Markov-chain algorithms for generating bipartite graphs and tournaments », in *Proceedings of the eighth annual ACM-SIAM symposium on Discrete algorithms*, SAKS Michael (dir.), Philadelphia, Society for Industrial and Applied Mathematics, 1997, p. 193-200, *repr. in Random Structures and Algorithms*, vol. 14, 1999, 293-308.

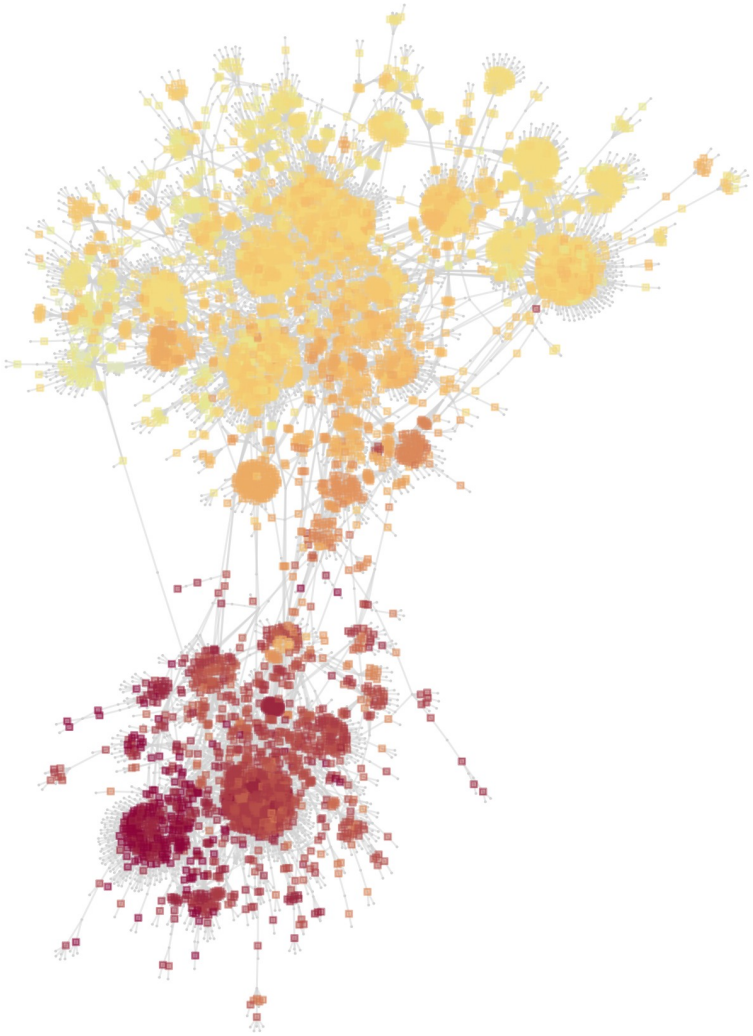


Figure 3: Plus grande composante connexe du graphe biparti : celle-ci est visualisée à l'aide d'un algorithme de forces (on considère que les sommets se repoussent mutuellement, sauf quand ils sont reliés par une arête, situation dans laquelle ils s'attirent au contraire). Les transactions sont représentées par des carrés et les individus par des petits cercles. Les couleurs, relatives aux transactions, donnent une information sur la date de la transaction (le rouge est utilisé pour les transactions les plus récentes et le jaune pour les transactions les plus anciennes).

Une fois la plus grande composante connexe extraite, celle-ci peut être visualisée en utilisant des logiciels spécifiques qui permettent des explorations interactives de graphes, comme par exemple le logiciel libre Gephi[10]. La Figure 3 fournit, par exemple, une représentation de la plus grande composante connexe du graphe obtenu à partir du corpus d'actes notariés. Celle-ci a été réalisée en utilisant un algorithme similaire à celui de Fruchterman et Reingold[11] mais incluant un processus de positionnement des sommets en deux étapes, comme décrit dans la thèse de doctorat de D. Tunkelang[12]. La figure ne permet pas de comprendre les détails locaux du graphe mais permet

10 Le logiciel est disponible a « Gephi – Makes Graphs Handy » http://www.gephi.org. Voir Bastian Mathieu, Heymann Sébastien, Jacomy Mathieu, « Gephi: an open source software for exploring and manipulating networks », in *Proceedings of the Third International AAAI Conference on Weblogs and Social Media*, Adar Eytan *et al.*, Menlo Park, AAAI Press, 2009, p. 361-62.

11 Fruchterman Thomas M.J., Reingold Edward M., « Graph drawing by force-directed placement », in *Software-Practice and Experience*, vol. 21, 1991, p. 1129-64.

12 Tunkelang Daniel, « A numerical optimization approach to general graph drawing ». Thèse de doctorat non publiée,

néanmoins de mettre en valeur des phénomènes globaux qui, par la visualisation, deviennent évidents. En particulier, la Figure 3 montre que le graphe est divisé en deux parties connexes faiblement connectées (la partie haute et la partie basse de la figure) qui sont elles-mêmes organisées en sous-structures denses, ce qui confirme ce que l'analyse de la connectivité du graphe, décrite plus haut, nous avait fait sentir : le graphe ne présente pas une répartition uniforme des arêtes mais admet, au contraire, des sous-structures denses faiblement connectées entre elles.

Pour aller plus loin, bien que la Figure 3 ait été réalisée sans utilisation a priori d'informations sur les dates des transactions (cette information a été ajoutée a posteriori, une fois la position des sommets déterminée par l'algorithme de forces, sous la forme d'une simple coloration), elle présente une organisation claire selon la date. Ainsi, la faible connectivité du graphe peut être expliquée par une analyse de la distribution des dates des transactions dans le corpus : cette distribution est représentée dans la Figure 4 (traits noirs). En particulier, on constate qu'aux environs de l'année 1400, le nombre d'actes notariés est très faible, ce qui est une conséquence directe de la peste noire et de la guerre de Cent Ans. Ainsi, la période la plus ancienne du graphe (la partie haute de la figure) et la période la plus récente (la partie basse de la figure) ne sont connectées que par les quelques transactions occasionnelles qui ont lieu autour de 1400. L'utilisation combinée de différentes approches, basées sur des points de vue différents et complémentaires des données du corpus initial, permet donc d'avoir une idée précise de caractéristiques macroscopiques importantes sur le corpus. Si la Figure 3 donne de manière immédiate une information sur la décomposition du graphe en sous-structures denses, cette conclusion pourrait être remise en cause sans les confirmations apportées à la fois par l'étude de la connectivité menée à l'aide d'une approche par simulation et par l'étude de la distribution des dates des transactions. À l'inverse, sans l'utilisation d'une représentation visuelle, l'identification de sous-structures denses, correspondant à des périodes d'activité accrue, n'aurait pas été aussi facile et intuitive.

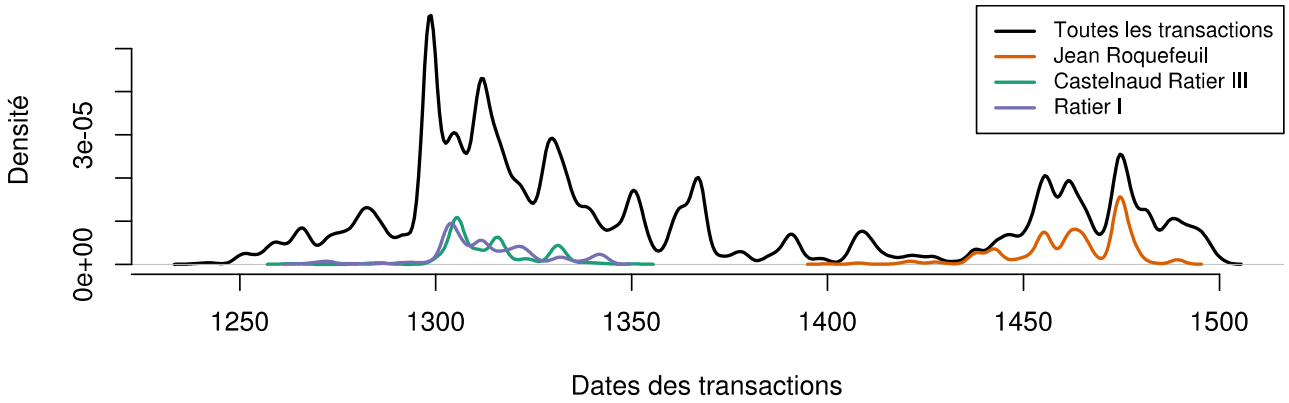


Figure 4: Distribution des dates des transactions dans le corpus et distribution des dates des transactions impliquant les trois principaux seigneurs.

Malgré tout, l'interprétation de la Figure 3 reste ardue et même en utilisant des outils d'exploration interactive de graphes, qui permettent des zooms et des déplacements manuels des sommets, l'utilisateur a de fortes chances de se retrouver totalement dépassé par la taille du graphe étudié. Une approche courante lorsqu'il s'agit d'appréhender des données de grande taille, comme celles-ci, est l'utilisation de méthodes de classification non supervisée qui permettent de regrouper les données en des sous-ensembles plus petits et relativement homogènes. Dans le cadre de l'analyse de graphe, la classification consiste à construire une partition des sommets d'un graphe en groupes denses de telle façon que les différents groupes soient relativement peu connectés[13]. La classification de sommets

School of Computer Science, Carnegie Mellon University, 1999, CMU-CS-98-189.

13 Deux articles récents proposent une synthèse sur les méthodes de classification dans les graphes : Fortunato Santo, « Community detection in graphs », in *Physics Reports*, vol. 486, 2010, p. 75-174 et Schaeffer Satu Elisa, « Graph clustering », in *Computer Science Review*, vol. 1, 2007, p. 27-64.

d'un graphe biparti reste encore un problème ouvert. Une solution consiste à construire indépendamment une classification des sommets des deux types (individus et transactions) tout en assurant la cohérence entre les deux classifications[14]. Néanmoins, nous montrerons plus tard que, dans le cas du présent graphe, les transactions et les individus ont des connectivités très différentes et qu'une telle approche pourrait amener un biais de classification. Aussi, nous avons réalisé une classification basée uniquement sur un graphe des individus[15], induit par le graphe biparti. Comme cela a été évoqué dans l'introduction, le graphe complet peut être *projeté* selon l'axe des individus ou selon celui des transactions. Ici, nous avons considéré le graphe dans lequel chaque sommet représente un individu du graphe initial (ou plutôt de sa plus grande composante connexe) ; dans ce graphe, deux sommets sont reliés par une arête si les deux individus correspondants sont actifs dans au moins une transaction commune. Cette approche fournit une classification des sommets en 34 classes dont la taille varie de 2 à 400 individus, avec un nombre moyen de personnes par classe égal à approximativement 110. Notons encore une fois qu'il ne s'agit pas d'inférer du graphe informationnel un réseau social sur les individus, mais, de façon beaucoup moins ambitieuse, de regrouper des individus d'une façon qui respecte le graphe informationnel et qui sera donc en cohérence avec la représentation visuelle du dit graphe, ceci dans l'optique de faciliter l'exploitation de cette représentation.

La Figure 5 montre ainsi comment une classification des sommets du graphe peut être utilisée pour améliorer la représentation qui en est donnée : cette visualisation améliorée met mieux en valeur les parties denses du graphe et leurs connexions respectives. Pour chaque classe, un individu *dominant* a été défini qui correspond à la personne ayant été active dans le plus grand nombre de transactions. Les 34 individus dominants ont été mis en valeur sur cette figure par leur nom et les classes ont été matérialisées par des cercles centrés sur l'individu dominant et dont la surface est proportionnelle au nombre d'individus de la classe. La connectivité entre les classes est représentée de manière simplifiée par des arêtes entre classes : l'épaisseur de ces arêtes est proportionnelle au nombre de transactions communes entre les membres des deux classes qu'elles relient. Ce résumé graphique simplifié apporte une perspective nouvelle sur le corpus documentaire, perspective plus facile à appréhender intuitivement que la représentation initiale du graphe. L'organisation du graphe en sous-structures denses y est mise en valeur et on voit clairement des concentrations d'arêtes épaisses entre certaines classes situées dans des zones différentes du graphe (par exemple au centre de la partie située en haut de la figure). Également, cette représentation met en valeur des liens spécifiques qui connectent certaines classes de la période la plus ancienne à des classes de la période la plus récente, par exemple entre les classes identifiées par Hélène Castelnau, Guy de Moynes et Arnaud Gasbert del Castanhier. Bien que ce résumé graphique ne remplace pas une analyse fine des détails du corpus, il permet d'avoir une carte globale du graphe, carte à laquelle l'utilisateur peut se référer lorsqu'il zoome sur des détails d'intérêt. Cette carte peut aussi guider l'utilisateur vers des problèmes dans les données saisies dans la base de données. En effet, on peut noter des occurrences multiples de noms identiques (par exemple, les deux classes nommées « Ratier » sur la partie en haut à droite de la figure) ou bien des connexions directes entre des classes dont les dates semblent très éloignées (par exemple, la connexion directe entre la classe nommée « Ratier » dans la partie située en haut à droite de la figure et la classe nommée « Jean Laperarede » dans la partie située en bas à gauche de la figure). Ce type de problèmes, qui sont pointés de manière très simple par une visualisation adaptée, peuvent ensuite être analysés plus en détail par des outils issus de la théorie des graphes. Si l'on s'intéresse de plus près au lien direct existant entre les classes nommées, respectivement, « Ratier » et « Jean Laperarede », on peut

14 Barber Michael J., « Modularity and community detection in bipartite networks », *Physical Review E*, vol. 76, 2007, p. 1-9.

15 La description précise de la méthode de classification utilisée peut être trouvée dans Rossi Fabrice, Villa-Vialaneix Nathalie, « Représentation hiérarchique d'un grand réseau à partir d'une classification hiérarchique de ses sommets », in *Journal de la Société Française de Statistique*, vol. 152, 2011, p. 34-65.

retrouver que ce lien correspond en fait à trois plus courts chemins[16] entre « Ratier » et « Jean Laperarede », chemins qui relient directement ces deux classes : ces trois chemins impliquent tous les trois les individus Bernard Garrigue, Arnaud Escairac, Berenguier Laperarede et une quatrième personne qui est différente dans chacun des trois chemins. Ainsi, l'analyse visuelle exploratoire permet d'identifier une liste de personnes à étudier plus en détail pour vérifier la correction des informations enregistrées dans la base de données. Nous poursuivrons cette analyse locale dans la section suivante, après avoir présenté d'autres méthodes automatiques pour extraire des personnes importantes ou intéressantes de la masse des individus présents dans le graphe.

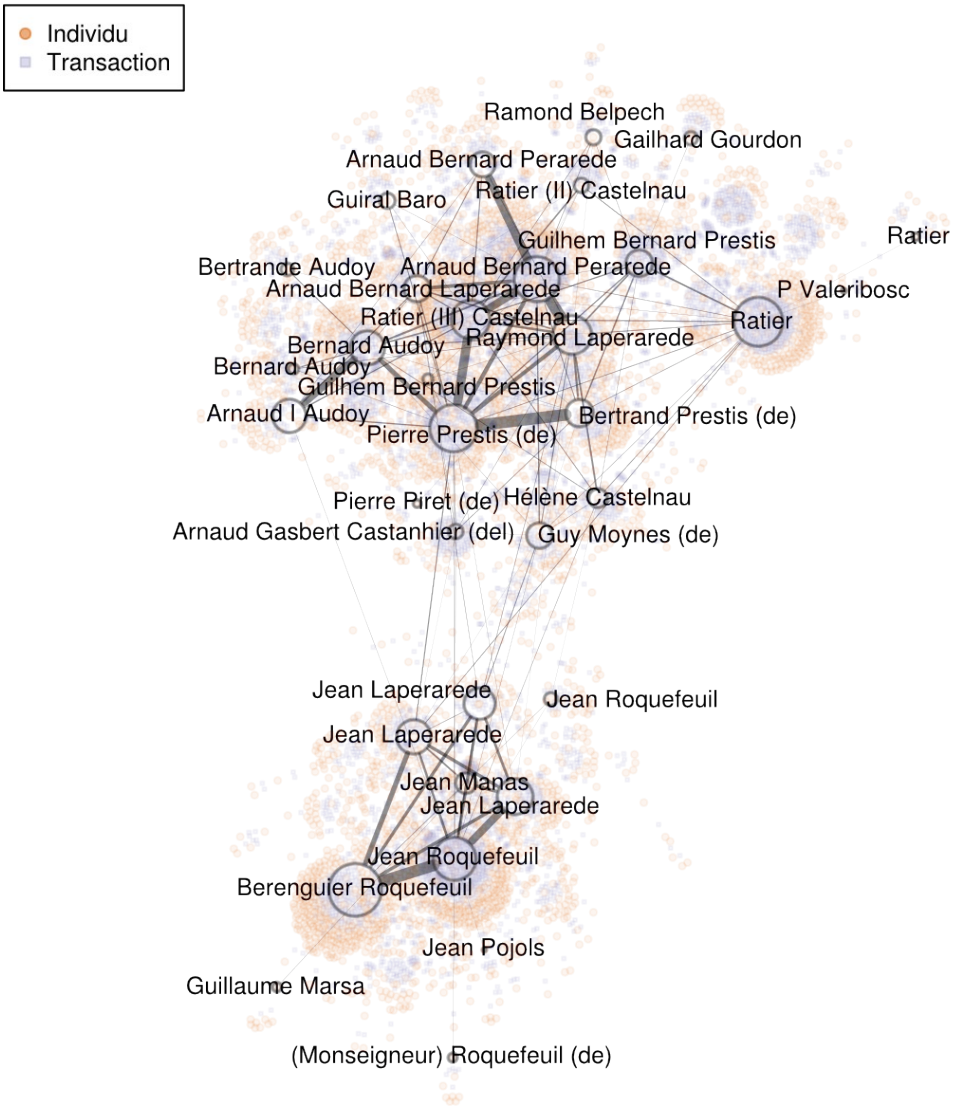


Figure 5: Plus grande composante connexe du graphe biparti. Une représentation simplifiée de la classification obtenue grâce au graphe projeté des individus a été ajoutée à cette figure : chaque cercle correspond à une classe qui a une surface proportionnelle à la taille de la classe qu'il représente. Les centres des classes sont positionnés sur l'individu ayant contracté le plus de transactions parmi les individus de la classe (que nous appelons individu « dominant » de la classe).

16 Le plus court chemin (ou parfois les plus courts chemins) entre deux sommets dans un graphe est un chemin reliant les deux sommets et dont le nombre d'arêtes est minimal.

Analyse locale

L'analyse globale du graphe, présentée dans la section précédente, a permis de donner une idée globale de la configuration du corpus documentaire avec des pistes pour détecter des sous-structures ou bien des erreurs de transcription des documents écrits dans la base de données numérique. Cependant, ces premières trouvailles doivent être confirmées par une analyse plus fine des phénomènes mis en valeur. Un des aspects qui ressort de l'analyse globale est la faible connectivité du graphe et sa structuration en sous-ensembles denses. Or, la connectivité du graphe peut aussi être analysée à l'échelle locale, par un changement de perspective, en déterminant des caractéristiques numériques propres aux sommets du graphe. Parmi les caractéristiques numériques disponibles pour les sommets de graphes, on peut s'intéresser, en particulier, au « degré » et à l'« intermédiarité ».

Le « degré » d'un sommet est le nombre d'arêtes afférentes à ce sommet. Le degré peut être interprété comme une mesure de la « popularité » d'un sommet. Dans le graphe biparti que nous étudions ici, il s'agit donc, pour un individu, du nombre de transactions dans lesquelles cet individu a été impliqué et, pour une transaction, du nombre d'individus impliqués dans celle-ci. On s'attend donc à ce que les degrés des individus et les degrés des transactions soient assez différents, les transactions impliquant typiquement un faible nombre d'individus alors que certains individus ont passé un nombre très importants de transactions. En fait, comme c'est souvent le cas dans les graphes réels, la distribution des degrés des individus suit une loi de puissance[17] : la plupart des individus ont un degré très faible (1 815 personnes ont un degré égal à 1, par exemple) alors qu'un faible nombre de sommets ont un degré très grand en comparaison (le cas le plus extrême étant Jean Roquefeuil qui a été actif dans 551 transactions). Les individus de plus fort degré sont tous des nobles et sont actifs en tant que « seigneur » dans la plupart des transactions dans lesquelles ils sont cités.

En tant que tel, le degré d'un individu n'est pas une caractéristique propre à la modélisation sous forme de graphe du corpus, puisqu'il est le simple reflet de l'activité individuelle d'une personne. Même si des informations intéressantes peuvent être tirées de l'étude de cette quantité (comme, par exemple, le fait que deux femmes, Lombarde Laperarede et Hélène Castelnau, fassent partie des 33 individus avec les plus forts degrés), nous nous concentrons dans la suite de cet article sur des analyses qui utilisent de manière plus directe le modèle graphique. Utilisant la classification décrite dans la section précédente, on peut comparer la liste des individus de plus fort degré avec la liste des individus dominants de chaque classe (qui correspondent donc à degrés maximaux locaux). Ces deux listes ont 31 individus en commun mais des différences notables existent entre ces deux méthodes d'extraction d'individus que l'on peut interpréter comme importants : par exemple, Raimond Perarede, qui est l'individu dont le degré est le sixième plus fort dans le graphe (il est impliqué dans 234 transactions), n'est pas dans la liste des individus dominants car il est classé dans le même groupe qu'Arnaud Bernard Perarede qui apparaît dans 304 transactions. En fait, ces deux seigneurs partagent 51 transactions. Ce phénomène est répété deux autres fois dans le graphe : entre Bernard Audoy et Jacmes Audoy (116 transactions communes) et entre Raymond Laperarede et Gausbert Lauriac (52 transactions communes). Ces divergences entre les deux listes mettent en valeur des phénomènes intéressants et pointent vers des relations familiales particulières dans le corpus : Bernard et Jacme Audoy étaient frères et avaient hérité des terres communes de leur père (aussi appelé Bernard Audoy) : les contrats qu'ils ont ensuite passés avec les tenants de ces terres ont donc tous été faits en commun (la famille Audoy est, par ailleurs, bien représentée dans la partie située en haut à gauche du graphe de la Figure 5). Les deux autres paires identifiées de cette manière (Raymond Laperarede et Gausbert Lauriac ainsi qu'Arnaud Bernard Perarede et Raimond

17 Barabási Albert-László,Albert Réka, « Emergence of scaling in random networks », in *Science*, vol. 286, 2009, p. 509-512.

Perarede) correspondent toutes les deux au même phénomène : ces couples sont impliqués dans un faible nombre d'actes (deux ou trois) mais qui incluent chacun un grand nombre de transactions (environ 33 pour un des actes au moins) : Arnaud Bernard Perarede et Raimond Perarede ont procédé à un échange de terres de grande envergure alors que Raymond Laperarede et Gausbert Lauriac ont procédé à une vente simultanée d'un grand nombre de terres. On peut imaginer que ces opérations ont eu des conséquences importantes sur l'organisation sociale locale.

À l'inverse, certains individus de faible degré se retrouvent parmi la liste des individus dominants dans la classification : treize classes ont en effet un individu dominant dont le degré est inférieur à 34 alors que le t34ème individu de plus fort degré a un degré égal à 53. Dans tous les cas, ces individus dominants sont associés à des classes de faible effectif qui doivent être examinées plus en détail car elles peuvent correspondre, là encore, à des erreurs de transcription. Par exemple, on retrouve deux « Guilhem Bernard Prestis » associés à deux classes distinctes : l'un est un individu de fort degré (impliqué dans 204 transactions) alors que l'autre est un individu de faible degré (impliqué dans seulement 12 transactions), ce qui est une indication que ces deux sommets pourraient ne former qu'une seule et unique personne.

Une autre caractéristique numérique standard dans l'analyse de graphe et particulièrement de réseaux sociaux, est l'« intermédiarité » d'un sommet qui correspond au nombre total de plus courts chemins entre les paires de sommets du graphe qui passent par celui-ci. Les sommets de forte intermédiarité sont donc ceux qui sont les plus susceptibles de casser la connexité du réseau si on les en enlève : c'est une mesure non locale, puisqu'elle utilise l'intégralité du graphe, mais c'est aussi une mesure spécifique de la théorie des graphes, qui n'existe pas en dehors d'une interprétation en réseau de la configuration informationnelle des données. Là encore, la liste des sommets de plus forte intermédiarité est très similaire à la liste des sommets de plus fort degré : parmi 34 individus, les deux listes possèdent 24 individus communs. Une attention particulière doit être portée aux individus de forte intermédiarité qui n'ont pas un fort degré : pour certains d'entre eux l'interprétation de cette particularité est aisée. Le « chapitre de Cahors » ou l' « église de Flaugnac » ont une large intermédiarité parce que ce sont des personnes morales, non mortelles. Ces noms apparaissent donc dans des transactions passées à des périodes très différentes et se retrouvent ainsi sur les plus courts chemins reliant les individus appartenant à ces différentes périodes. Pour les autres, une analyse plus précise doit être menée, utilisant différents aspects du corpus, et combinant en particulier le modèle graphique et des informations associées à seulement un des deux types de sommets.

Propager des informations

Les modèles relationnels sont couramment utilisés pour propager des informations : de manière tout à fait intuitive, si une transaction est caractérisée par une date, un lieu, elle fournit des informations sur la période d'activité ou la zone d'activité des individus qui y sont impliqués, même si le modèle relationnel de la base de données associe ces informations à la transaction et non aux individus (pour des raisons évidentes de bonne conception et de mise sous forme normale de la base de données). Ce faisant, on peut voir que les individus avec une large intermédiarité et un faible degré, mis en exergue dans la section précédente, sont des individus dont la période d'activité est anormalement grande. C'est le cas du chapitre de Cahors dont la période d'activité va de 1277 à 1472. Mais si ce fait est compréhensible pour une personne morale, il devient suspicieux lorsqu'il s'agit d'un être humain avec une durée de vie irréaliste. Par exemple, « Arnaud Escairac » apparaît dans seulement dix transactions mais celles-ci sont datées de 1333 à 1481 : ceci est clairement une erreur de transcription et sous le nom de « Arnaud Escairac » se cachent en fait probablement plusieurs homonymes. La Figure 6 fournit une description plus précise de ce problème : « Arnaud Escairac » est impliqué principalement dans des transactions datées autour de 1479 avec des

personnes qui sont mentionnées dans d'autres transactions de dates similaires. Par ailleurs, il apparaît dans deux transactions datées de 1333 qui impliquent des personnes elles-mêmes impliquées dans des transactions de dates similaires. Ainsi, les deux sous-graphes (celui correspondant aux transactions datées des environs de 1333 et celui correspondant aux transactions datées des environ de 1479) semblent être tous les deux cohérents et doivent correspondre à deux homonymes vivant à des périodes différentes.

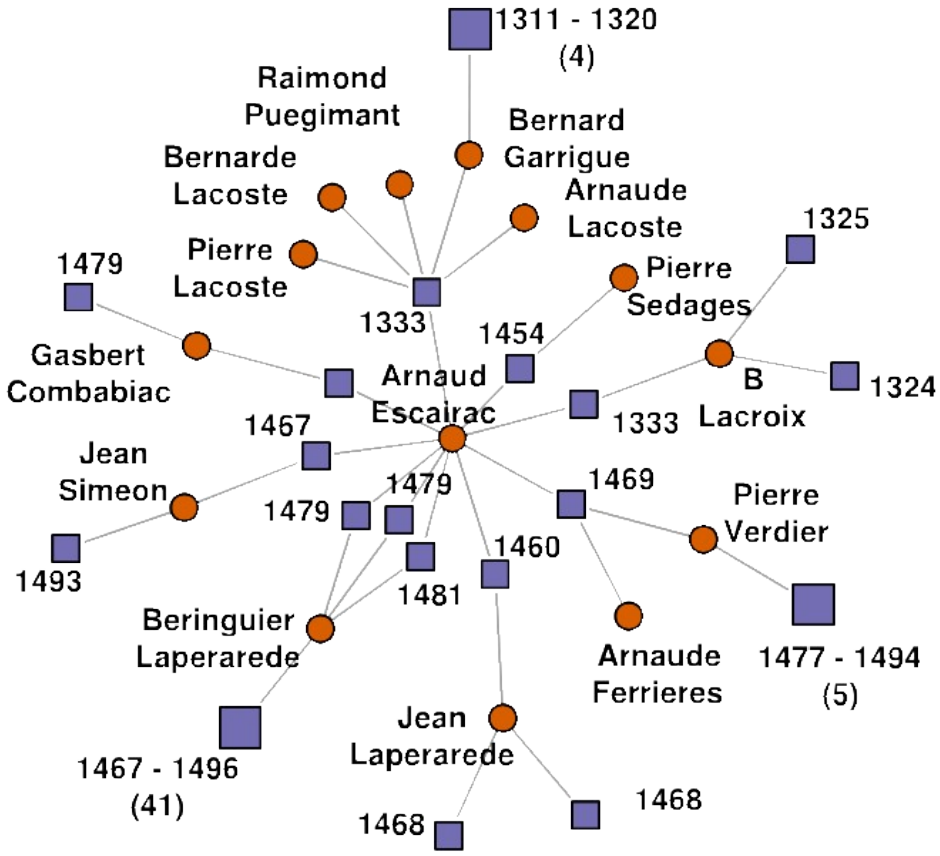


Figure 6: Réseau local autour de « Arnaud Escairac » jusqu'au troisième voisin. Les petits carrés correspondent à des transactions, les deux carrés plus étendus résument deux ensembles de transactions (41 en bas à gauche et 5 en bas à droite) et les cercles correspondent aux individus. Les dates des transactions sont inscrites sur les sommets correspondants, lorsqu'elles sont connues et, pour les deux groupes de transactions, les dates sont résumées par un intervalle.

Il est également intéressant de noter que « Arnaud Escairac » avait déjà été mis en lumière lors de l'analyse globale du réseau : il est apparu sur tous les plus courts chemins reliant Ratier à Jean Laperarede (voir le dernier paragraphe de la section Analyse macroscopique). C'est en fait cette erreur de transcription qui explique la connexion directe entre les classes dont Ratier et Jean Laperarede sont les individus dominants. Ceci explique aussi la valeur élevée de l'intermédiarité de « Arnaud Escairac » puisqu'à cause de cette erreur, quelques individus qui ont été classés dans la classe de Jean Laperarede auraient dû être classés dans celle de Ratier, compte tenu de leurs dates d'activité.

De tels cas sont, en fait, simples à détecter : la durée de vie d'« Arnaud Escairac » étant en elle-même suspicieuse, le graphe n'est qu'un moyen pratique de représenter le problème et de proposer une méthode simple pour désambiguïser facilement la base de données. Cependant, le même principe de propagation d'informations peut être mis en œuvre pour détecter et expliquer des situations moins évidentes. C'est par exemple le cas de « Guiral Combe », qui a une intermédiarité

élevée bien que n'apparaissant ni dans la liste des individus de plus fort degré, ni dans la liste des individus dominants une classe. Sa durée de vie estimée par propagation d'informations s'étend de 1318 à 1370, ce qui est élevé mais pas totalement impossible : dans ce cas, la configuration en réseau et plus particulièrement un zoom local autour de ce sommet, va prendre tout son sens pour fournir une explication possible.

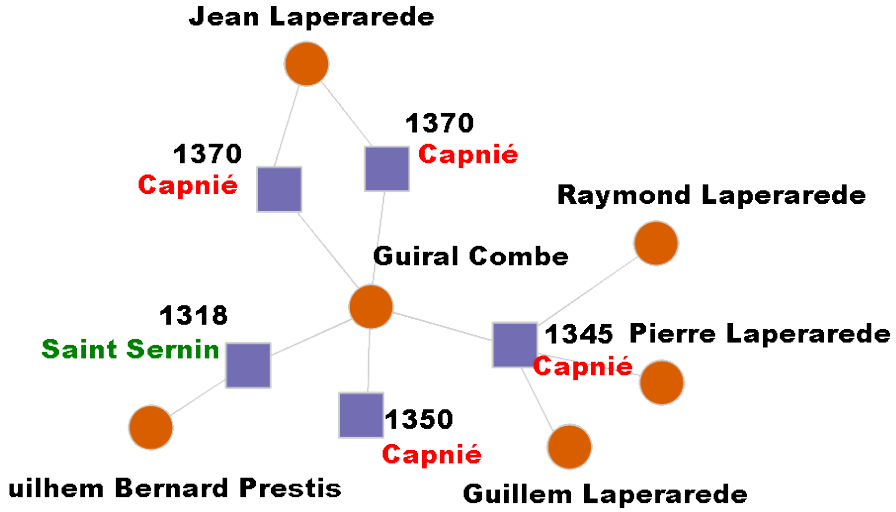


Figure 7: Voisinage local autour de « Guiral Combe » jusqu'au second voisin. Les carrés correspondent aux transactions et les cercles aux individus. Les textes désignant les paroisses associées aux transactions sont de deux couleurs différentes pour mieux différencier les dites paroisses. Les dates des transactions sont données à côté des transactions.

La Figure 7 représente le voisinage local de « Guiral Combe » dans la configuration en réseau du corpus documentaire. Le voisinage est représenté jusqu'à l'ordre deux, c'est-à-dire que seuls les sommets pouvant être atteints à partir de « Guiral Combe » par deux arêtes sont représentés. Des informations supplémentaires sont superposées sur la représentation du réseau : pour les sommets correspondant à des individus, leurs noms et, pour les sommets correspondant à des arêtes, leurs dates et les paroisses concernées. Avec ces informations supplémentaires, deux groupes distincts de sommets se détachent : le premier contient quatre transactions, toutes en rapport avec la paroisse de Capnié et avec des individus de la famille Laperarede. Ces transactions ont été établies entre 1345 et 1370. Le deuxième groupe contient une unique transaction, passée en 1318 avec Guilhem Bernard de Prestis au sujet d'un lieu situé dans la paroisse de Saint-Sernin. Cette information, couplée au fait que les paroisses de Saint-Sernin et Capnié ne sont pas limitrophes, est un indice fort en faveur d'une homonymie pour le sommet « Guiral Combe ». Cette hypothèse demandera à être confirmée par un retour aux sources historiques mais montre néanmoins comment une analyse utilisant la combinaison d'une interprétation relationnelle de la base documentaire et les informations disponibles sur les entités formant le réseau permet de repérer de manière automatique des erreurs de transcription qui auraient pu être difficiles à trouver directement à partir d'une vision tabulaire classique de la base de données.

Conclusion

Cet article a présenté comment une interprétation littérale de la structure relationnelle d'une base de données documentaire conduit à considérer un corpus comme un réseau auquel on peut appliquer des techniques classiques d'analyse de graphe et bénéficier ainsi d'outils d'analyse avancés pour aider à la compréhension de phénomènes globaux et locaux. Les modèles et l'approche

présentés ici sont suffisamment généraux pour pouvoir être appliqués dans une variété de problèmes du même type, à partir du moment où des entités d'intérêt sont liées par un ou plusieurs types de relations. L'ajout d'informations supplémentaires pour qualifier les sommets d'une configuration en réseau pourrait être complété par l'ajout d'informations qualifiant les arêtes (rôle dans la transaction) pour proposer une représentation encore plus précise du corpus initial.

La méthodologie présentée repose sur des outils de fouille de données, tirant le meilleur parti de la configuration étudiée du corpus, parmi lesquels on peut citer les méthodes de visualisation, de classification, le calcul d'indices numériques ou l'extraction de réseaux locaux. La visualisation et la classification ont pour but de fournir des représentations pertinentes, simplifiées et intuitives de la base de données pour aider l'œil humain à en comprendre sa structure macroscopique. Les indices numériques ou l'extraction de réseaux locaux peuvent, quant à eux, être utilisés pour trouver les individus importants ou atypiques et fournir des moyens automatiques pour détecter les anomalies et les erreurs de transcription qui auraient été difficiles à identifier directement. L'implémentation de telles approches dans des programmes de visualisation interactive de graphes, comme Gephi, permettrait de proposer aux chercheurs ne travaillant pas directement dans le domaine de l'informatique et des mathématiques, un moyen simple pour obtenir et exploiter, sur leurs données , les divers points de vue présentés dans cet article.