Approche réseau pour la fouille d'un grand corpus d'actes notariés médiévaux (F Hautefeuille, F. Rossi)

Fabrice Rossi

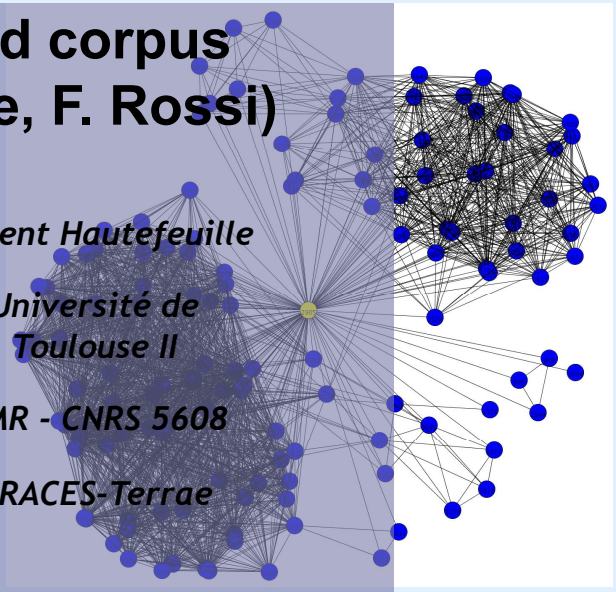
SAMM (Université Paris 1)

Florent Hautefeuille

Université de Toulouse II

UMR - CNRS 5608

TRACES-Terrae



avec Nathalie Villa (Toulouse 3, UMR 5219 IMT)

Traitement des données en SHS (difficultés et perspectives) : séminaire de la SFDS, 6 avril 2012

0 – Aux origines : le projet Graph-comp

Projet ANR dirigé par B. Jouve regroupant 6 UMR

Un pôle mathématique

-Institut de Mathématique de Toulouse

Un pôle informatique

-Institut de Recherche en Informatique de Toulouse (IRIT)

-LINA (école polytech de Nantes)

Un pôle histoire et archéologie

-FRAMESPA (Toulouse)

-TRACES (Toulouse)

0 – Aux origines : le projet Graph-comp

Projet ANR dirigé par B. Jouve regroupant 6 UMR

Un pôle mathématique

-Institut de Mathématique de Toulouse

Un pôle informatique

-Institut de Recherche en Informatique de Toulouse (IRIT)

-LINA (école polytech de Nantes)

Un pôle histoire et archéologie

-FRAMESPA (Toulouse)

-TRACES (Toulouse)

Objectifs = mettre en œuvre une approche mathématique pour l'étude des structures de la société rurale médiévale

1.1 – Une zone laboratoire :

La châtellenie de Castelnau-Montratier

- Bas-Quercy

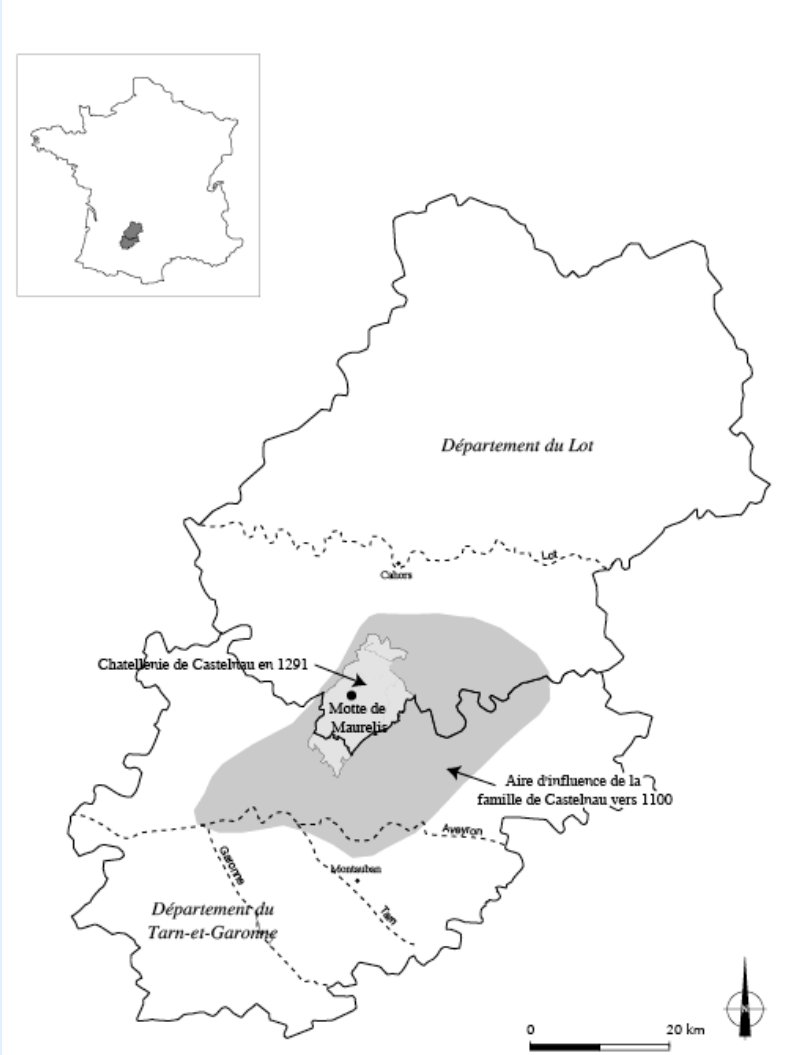


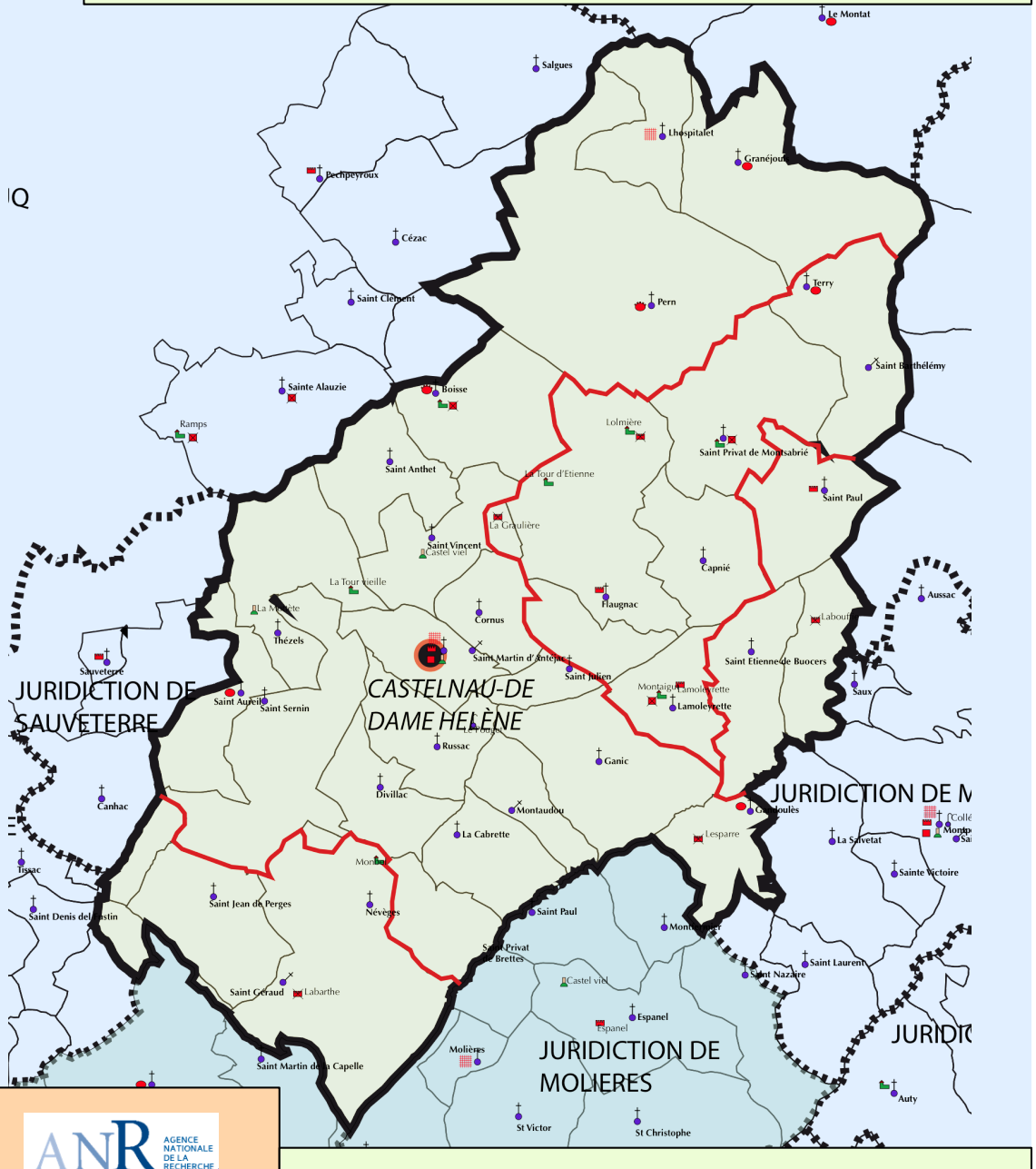
Fig. 1 - Localisation générale

1.1 – Une zone laboratoire :

La châtellenie de Castelnau-Montratier

- Bas-Quercy

- territoire d'une quarantaine de paroisses (± 300 km2)



Florent Hautefeuille, UT2 / Fabrice Rossi

1.1 – Une zone laboratoire :

La châtellenie de Castelnau-Montratier

-territoire d'une quarantaine de paroisses

-pays de Vaux = variété de paysages et des formes de peuplement

→ ± 15 villages vers 1300

-groupe de 10 à 100 maisons

-castrum, villa, bastide

Florent Hautefeuille, UT2 / Fabrice Rossi

1.1 – Une zone laboratoire :

La châtellenie de Castelnau-Montratier

-territoire d'une quarantaine de paroisses

-pays de Vaux = variété de paysages et

des formes de peuplement

→ 6 disparitions de village entre 1350 et 1700

Florent Hautefeuille, UT2 / Fabrice Rossi

1.1 – Une zone laboratoire :

La châtellenie de Castelnau-Montratier

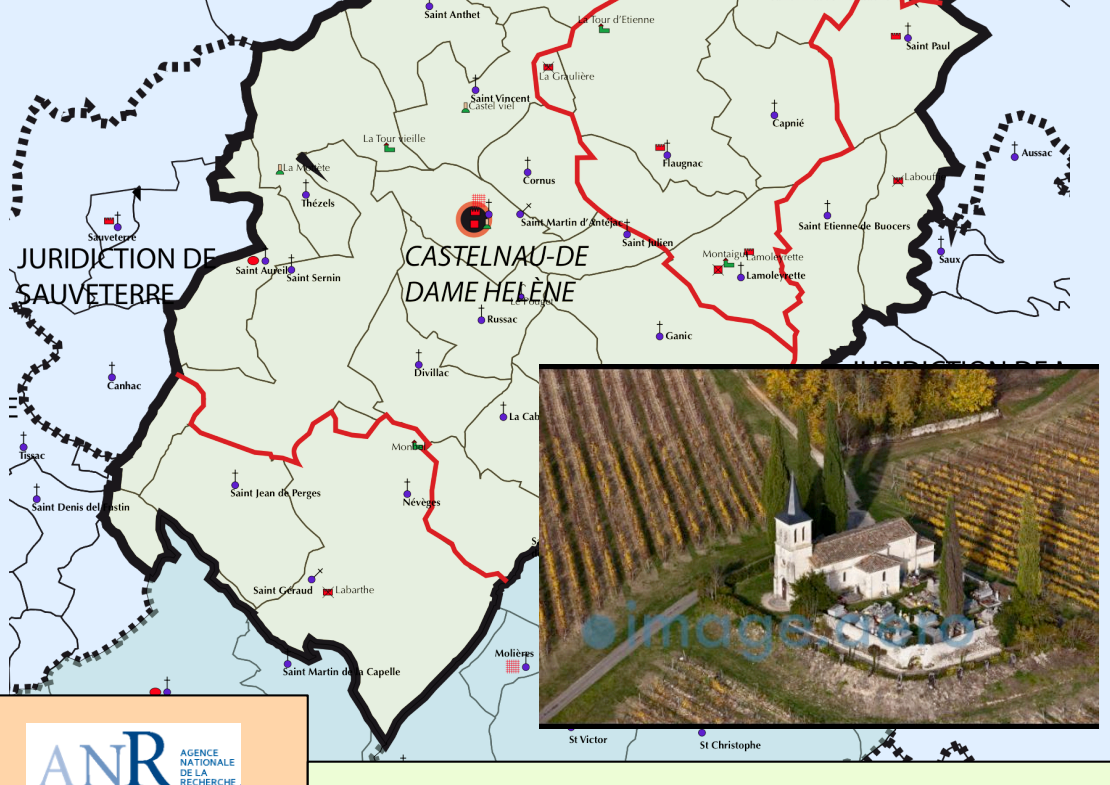
-territoire d'une quarantaine de paroisses

-pays de Vaux = variété de paysages et des formes de peuplement

→ une trame de peuplement fortement dispersé

-mas et cammas

-bordes

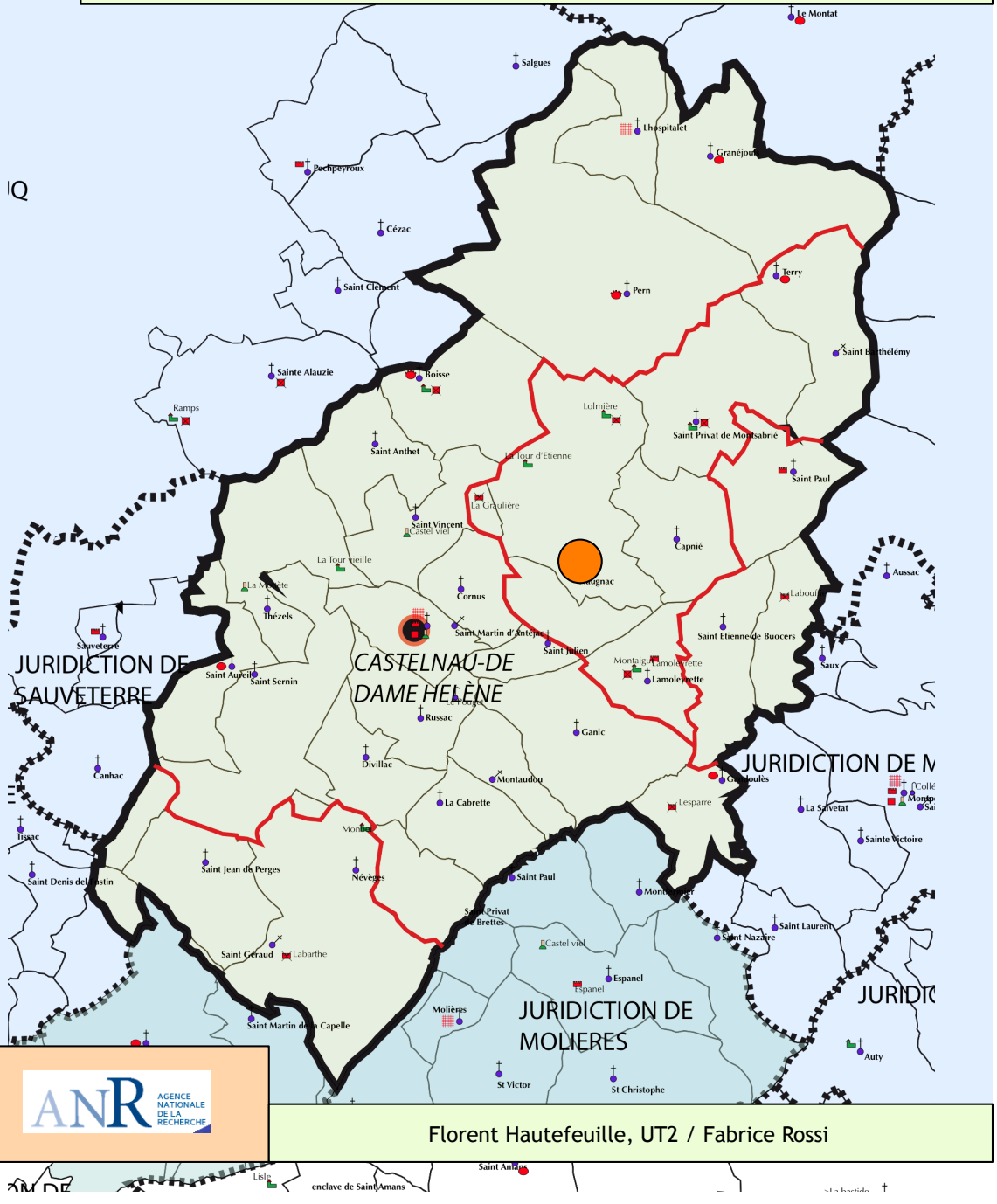


Florent Hautefeuille, UT2 / Fabrice Rossi

1.2 – Sources disponibles:

-les données archéologiques

-fouille du castrum de Flaugnac (2001-2003)



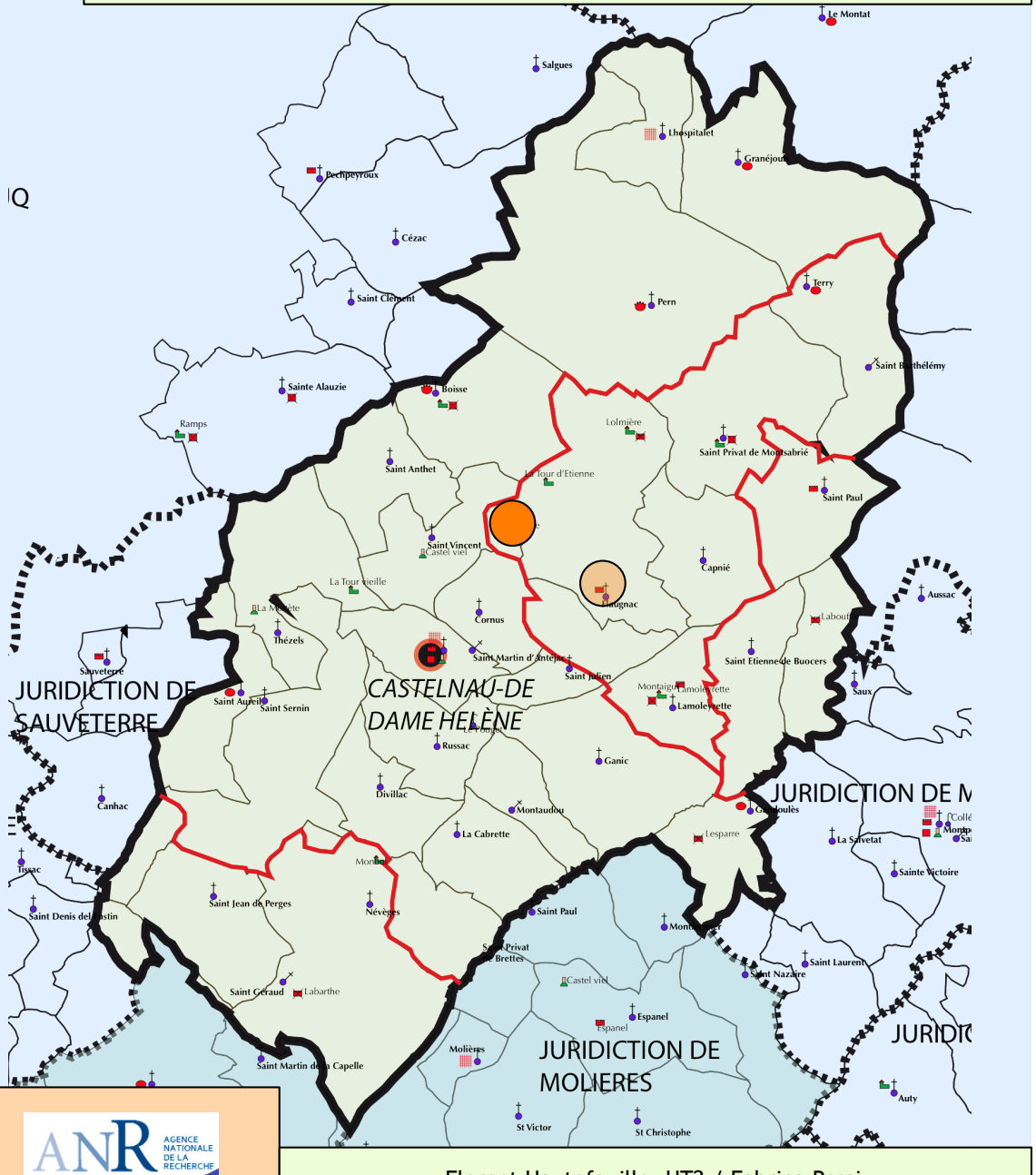
Florent Hautefeuille, UT2 / Fabrice Rossi

1.2 – Sources disponibles:

-les données archéologiques

-fouille du castrum de Flaugnac (2001-2003)

-fouille de la villa de la Graulière (2003-2004)



Florent Hautefeuille, UT2 / Fabrice Rossi

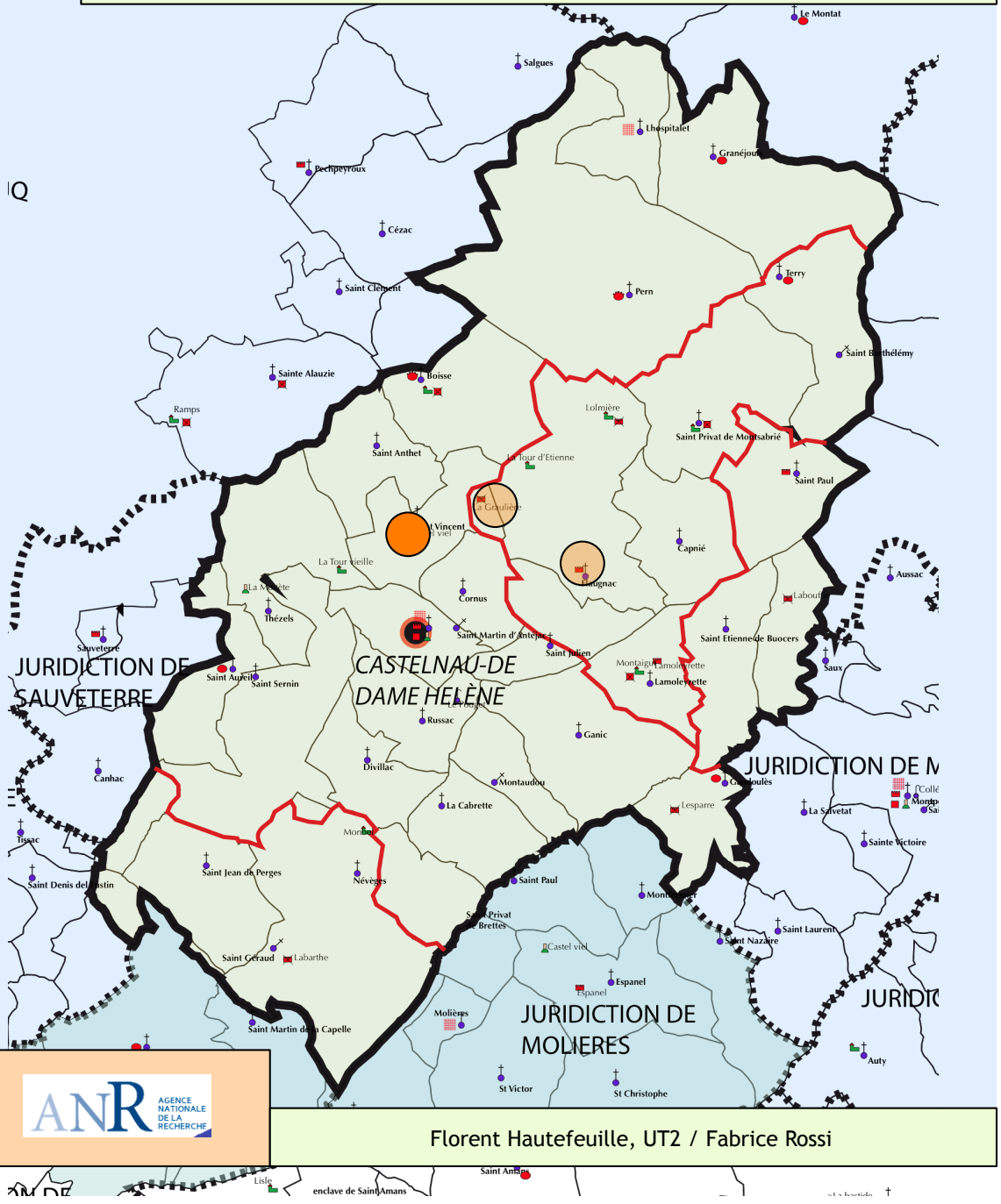
1.2 – Sources disponibles:

-les données archéologiques

-fouille du castrum de Flaugnac (2001-2003)

-fouille de la villa de la Graulière (2003-2004)

-fouille du castrum carolingien de Castel-viel et de la ferme tardo-médiévale associée (2004-2008)



1.2 – Sources disponibles:

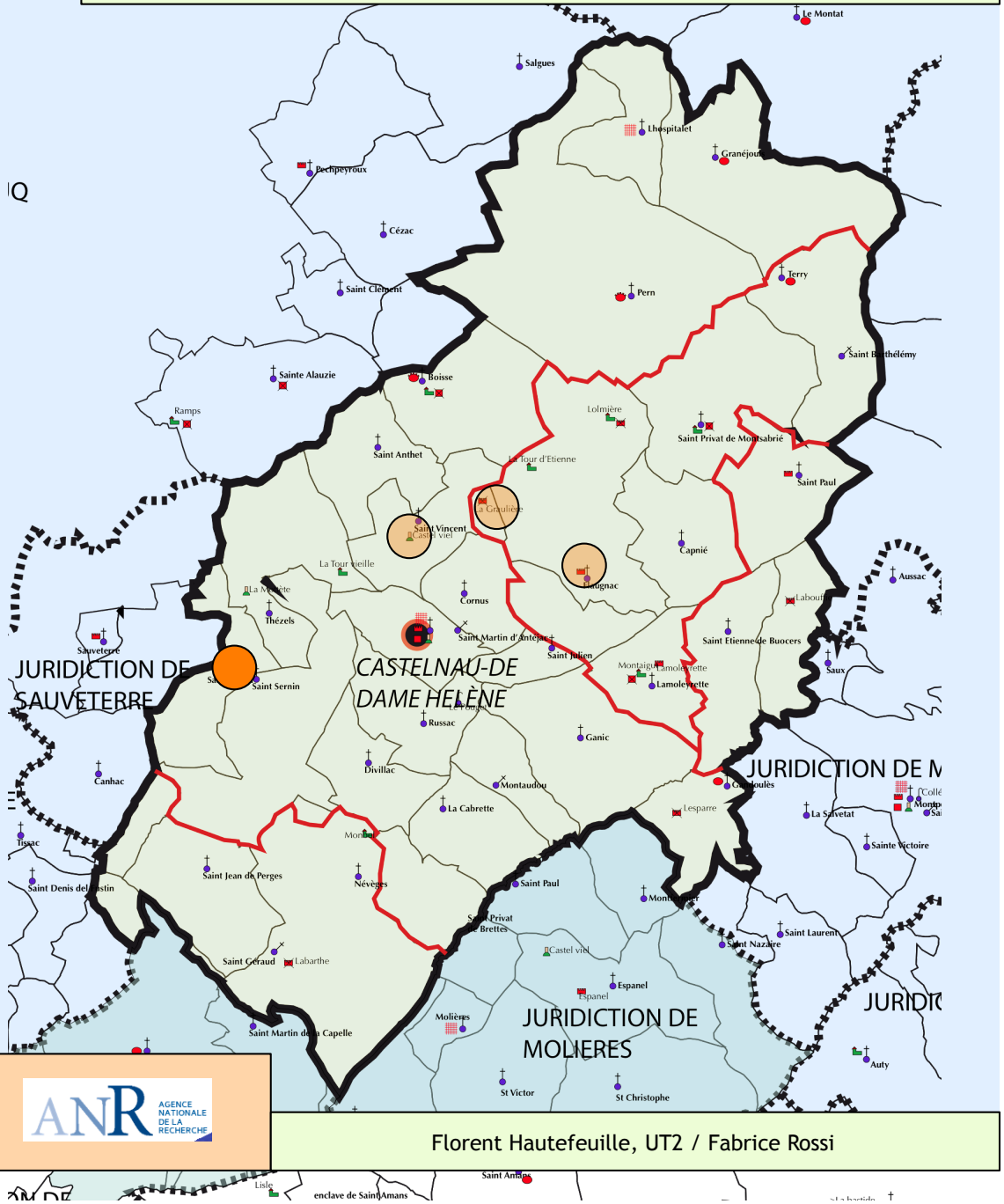
-les données archéologiques

-fouille du castrum de Flaugnac (2001-2003)

-fouille de la villa de la Graulière (2003-2004)

-fouille du castrum carolingien de Castel-viel et de la ferme tardo-médiévale associée (2004-2008)

-étude archéologique et historique de l'église emmottée et du castrum de Saint Aureil (2009-2011)



Florent Hautefeuille, UT2 / Fabrice Rossi

1.2 – Sources disponibles:

-les données archéologiques

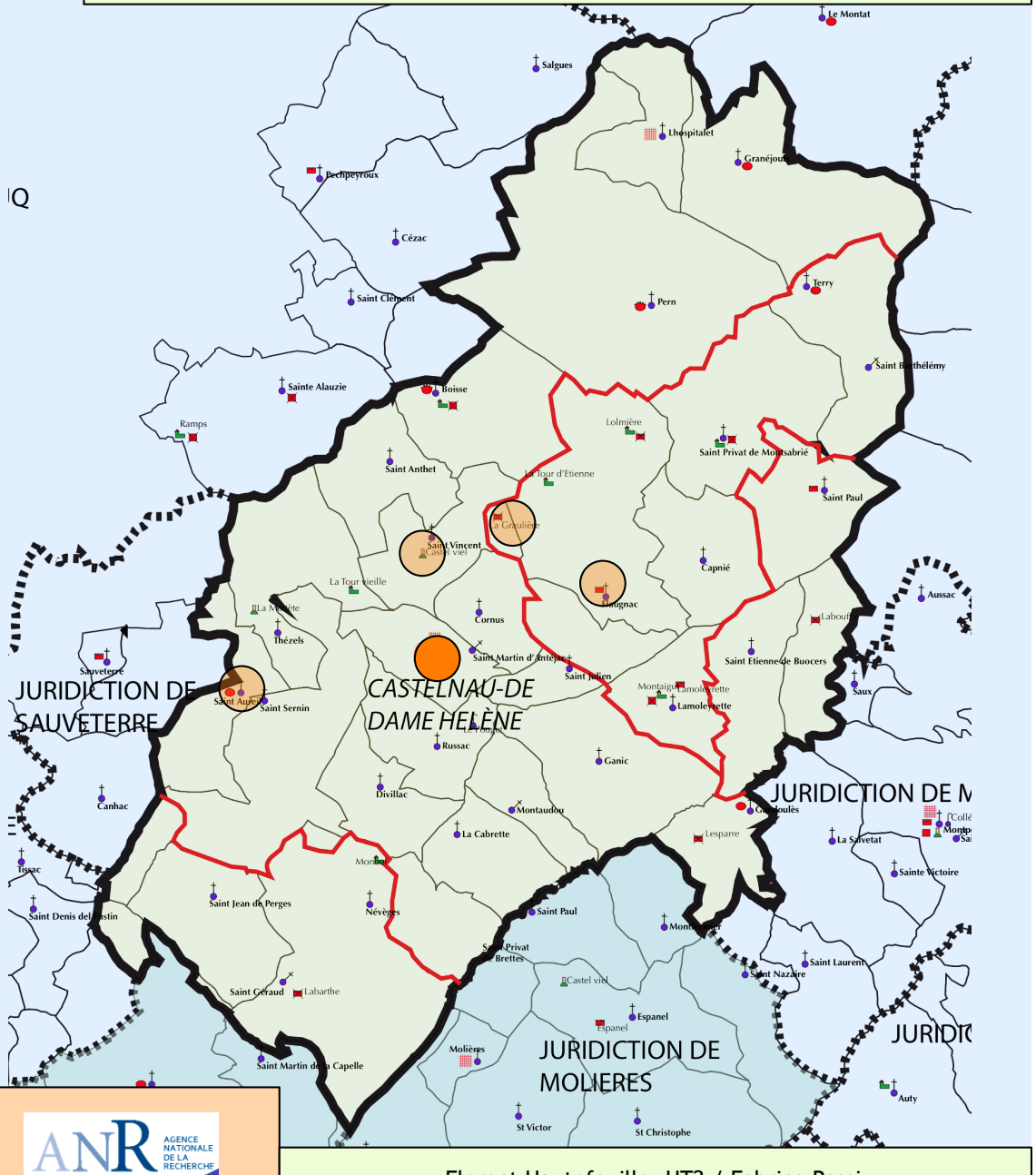
-fouille du castrum de Flaugnac (2001-2003)

-fouille de la villa de la Graulière (2003-2004)

-fouille du castrum carolingien de Castel-viel et de la ferme tardo-médiévale associée (2004-2008)

-étude archéologique et historique de l'église emmottée et du castrum de Saint Aureil (2009-2011)

-étude morphologique du bourg de Castelnau



Florent Hautefeuille, UT2 / Fabrice Rossi

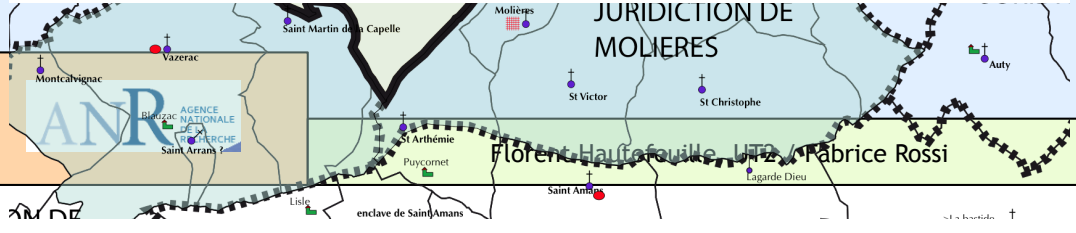
1.2 – Sources disponibles:

-les données archéologiques

-identification nominative de la plupart de structures étudiées

ex. les deux bâtiments fouillés sur la villa de la Graulière appartiennent à Jean Don et Bernard Teichendier, très largement documentés dans la BDD

La *villa* de La Graulière vers 1300



Florent Hautefeuille, UTM / Fabrice Rossi

1.2 – Sources disponibles:

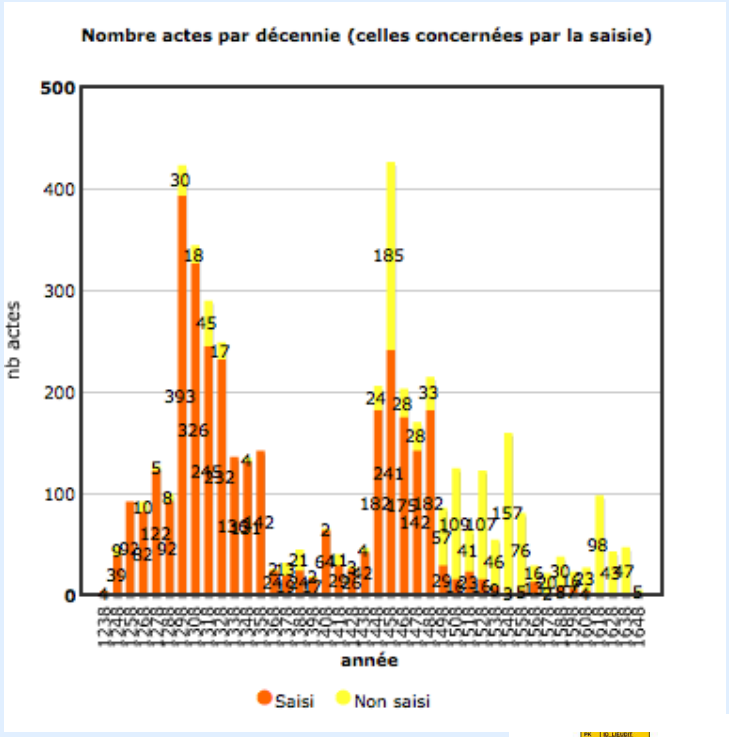
-les sources écrites

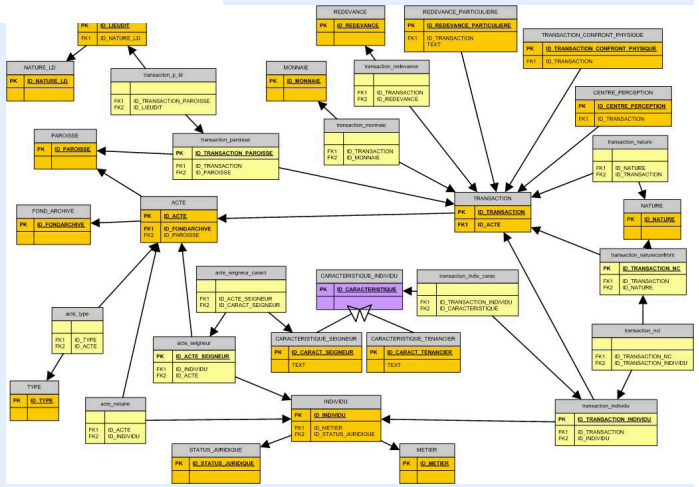
+ de 3350 actes notariés (essentiellement des contrats agraires) mis en base de données

+ de 8500 individus recensés et fichés

Période couverte = 1250-1500 avec deux grands périodes très fortement documentées (1250-1370 et 1440-1500) et une documentation plus faible entre les deux

Une base consultable : http://graphcomp.univ-tlse2.fr





Florent Hautefeuille, UT2 / Fabrice Rossi

1.2 – Sources disponibles:

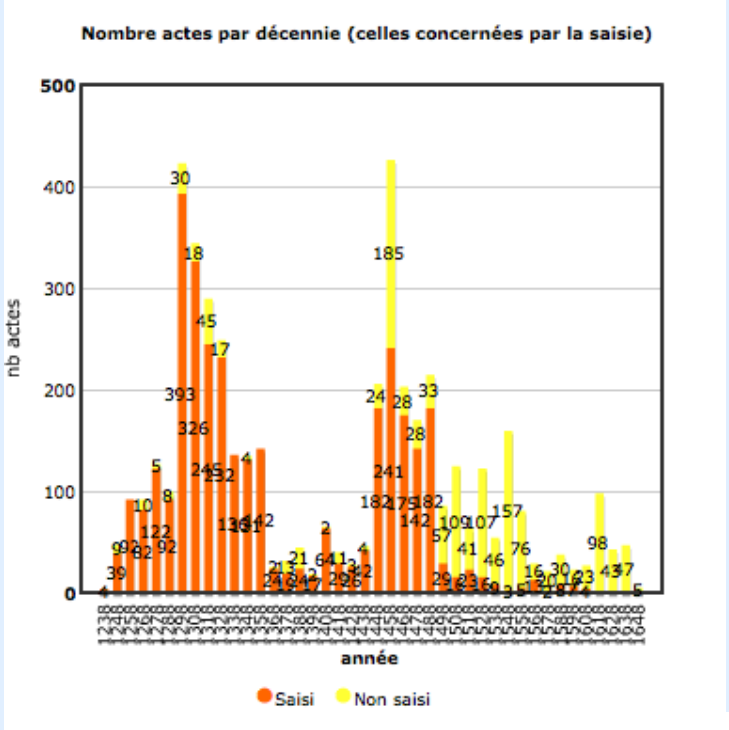
-les sources écrites

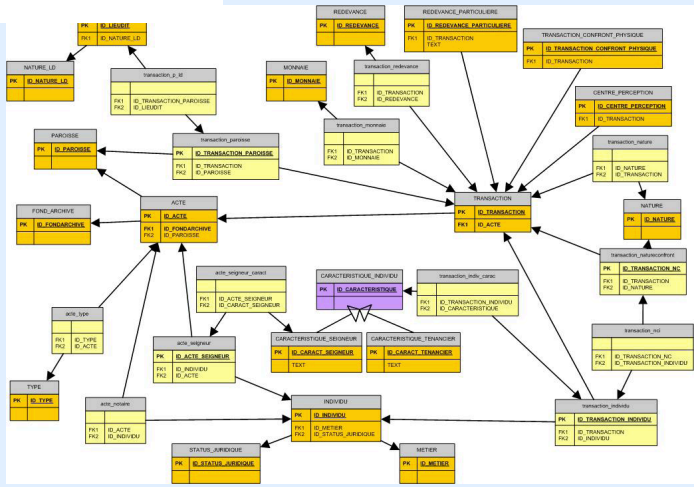
→ Difficultés inhérentes à ce corpus.

-problèmes de lecture des sources (saisie par plusieurs contrats de M-2 et doctorants)

-problèmes de la désambiguisation des noms de personne très imparfaite

-problèmes liés au faible nombre d'originaux et à un travail sur copies ou analyses détaillées des XVII° et XVIII°s.





Florent Hautefeuille, UT2 / Fabrice Rossi

1.2 – Les sources : écrites

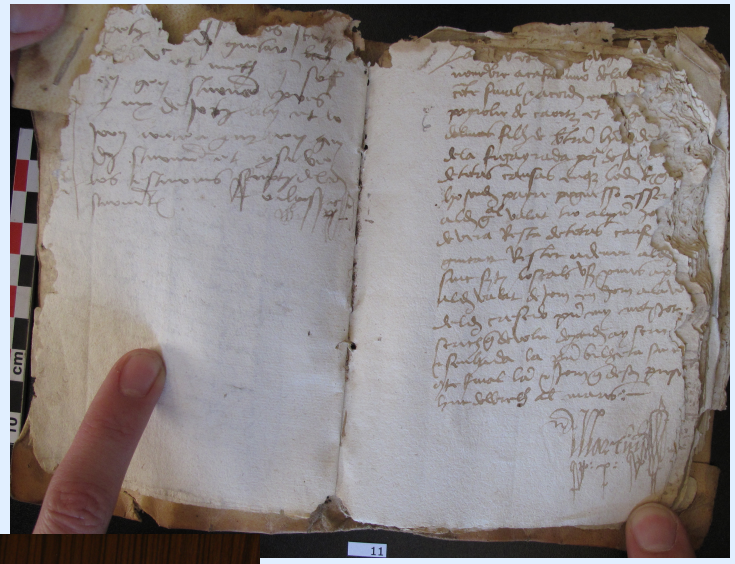
Des sources complémentaires :

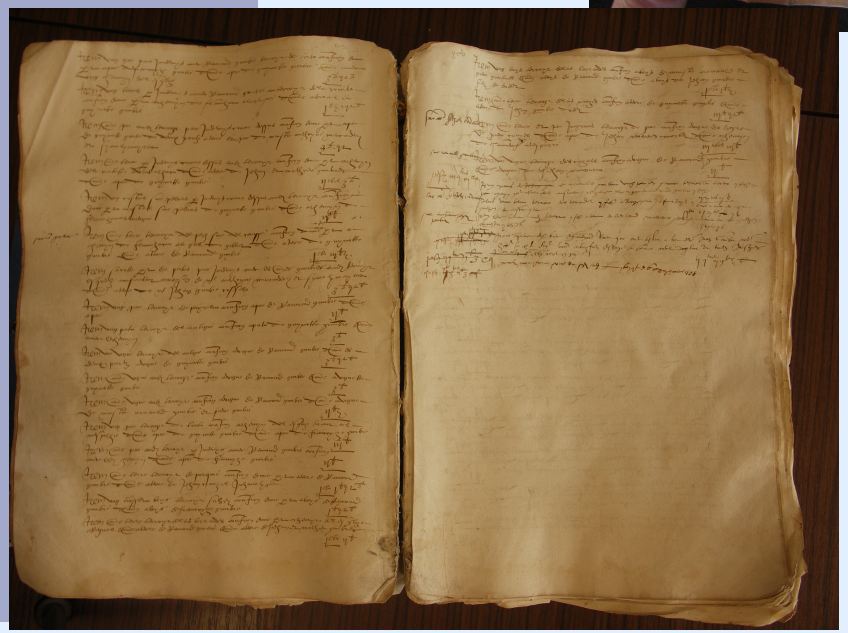
-Le livre de raison de la famille Guitard (1417-1526)

-Les compoix de 1530

-Un terrier de 1337

-Des terriers modernes





1.2 – Les sources :

Méthode = construction d'un graphe d'adjacence

➔Objectif = tenter une proche globale de cette masse documentaire

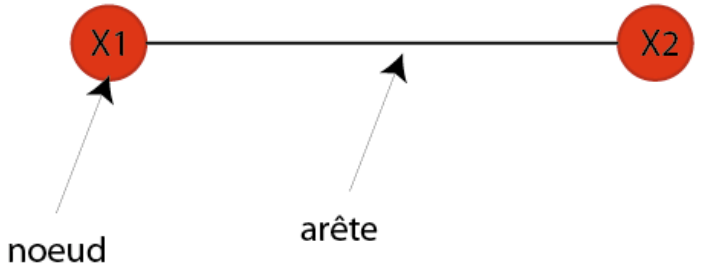
➔Passer par la construction des graphes des personnes présentes dans le corpus

-chaque individu est un noeud du graphe

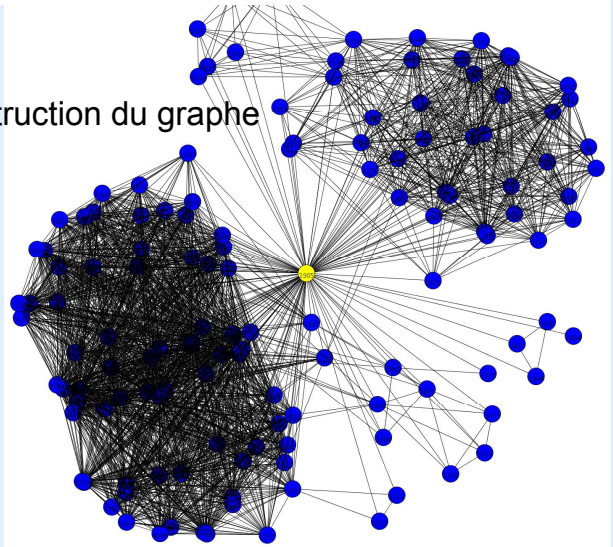
-chaque lien social entre individus se traduit par une arête du graphe

=> possibilité de travailler sur l'ensemble du corpus avec une vision globale, mais aussi de focaliser à tout moment l'analyse sur un individu ou un groupe d'individus

exemple : un individu X1 vend une parcelle à un individu X2



Modalités de construction du graphe



Florent Hautefeuille, UT2 / Fabrice Rossi

1.3 – Questionnement

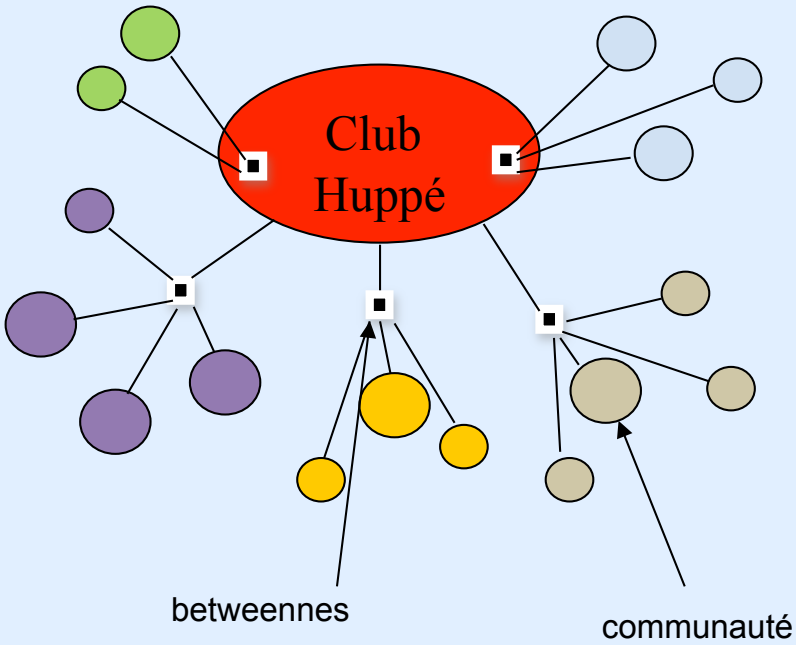
problème initial = comment analyser l'organisation globale de la société rurale de la fin du Moyen Age

➔ thèse de Romain Boulet

Boulet R. *Comparaison de graphes, application aux réseaux de sociabilités paysans du moyen âge", Doctorat de Mathématique, sous la dir. de B. Jouve, UT2, 2008.*

-mise en évidence d'un club huppé et d'individus relais

-déconnexion entre cette organisation et le structures sociales par ailleurs connues, comme le statut juridique (serf/libre, tenancier/seigneur)



1.3 – Questionnement

Objectifs actuels pour aller au-delà de ce premier niveau de résultat

→ caractériser l'évolution de ces structures à travers les 250 ans couverts par la documentation

-repérer l'impact des évolutions sociales fortes telles que la disparition du servage vers 1320-1350 ou la crise démographique de 1350-1380

1.3 – Questionnement

Objectifs actuels pour aller au-delà de ce premier niveau de résultat

→ caractériser l'évolution de ces structures à travers les 250 ans couverts par la documentation

-repérer l'impact des évolutions sociales fortes telles que la disparition du servage vers 1320-1350 ou la crise démographique de 1350-1380

-mieux comprendre l'impact du repeuplement de 1440-1480 sur le maintien ou l'évolution de cette organisation

→ dynamique temporelle des graphes de réseaux sociaux

1.3 – Questionnement

Objectifs actuels pour aller au-delà de ce premier niveau de résultat

→ prendre en compte des facteurs jusqu'alors peu exploités

-prise en compte de la nature du lien qui relie un individu à un autre (coloriage des arêtes). Un simple lien de voisinage est a priori moins fort qu'un contrat de mariage ou la vente de parcelles de terre

1.3 – Questionnement

Objectifs actuels pour aller au-delà de ce premier niveau de résultat

→ prendre en compte des facteurs jusqu'alors peu exploités

-prise en compte de la nature du lien qui relie un individu à un autre (coloriage des arêtes). Un simple lien de voisinage est a priori moins fort qu'un contrat de mariage ou la vente de parcelles de terre

-prise en compte des données spatiales. Chaque individu a un rattachement spatial (résidence, espace agricole, espace religieux...). Comment prendre en compte cet aspect de la BDD ?

→ dynamique spatiale des graphes de réseaux sociaux

1.3 – Questionnement

Objectifs actuels pour aller au-delà de ce premier niveau de résultat

→ limiter les effets déformants liés à la nature imparfaite de la source et de la BDD

-désambuigisation

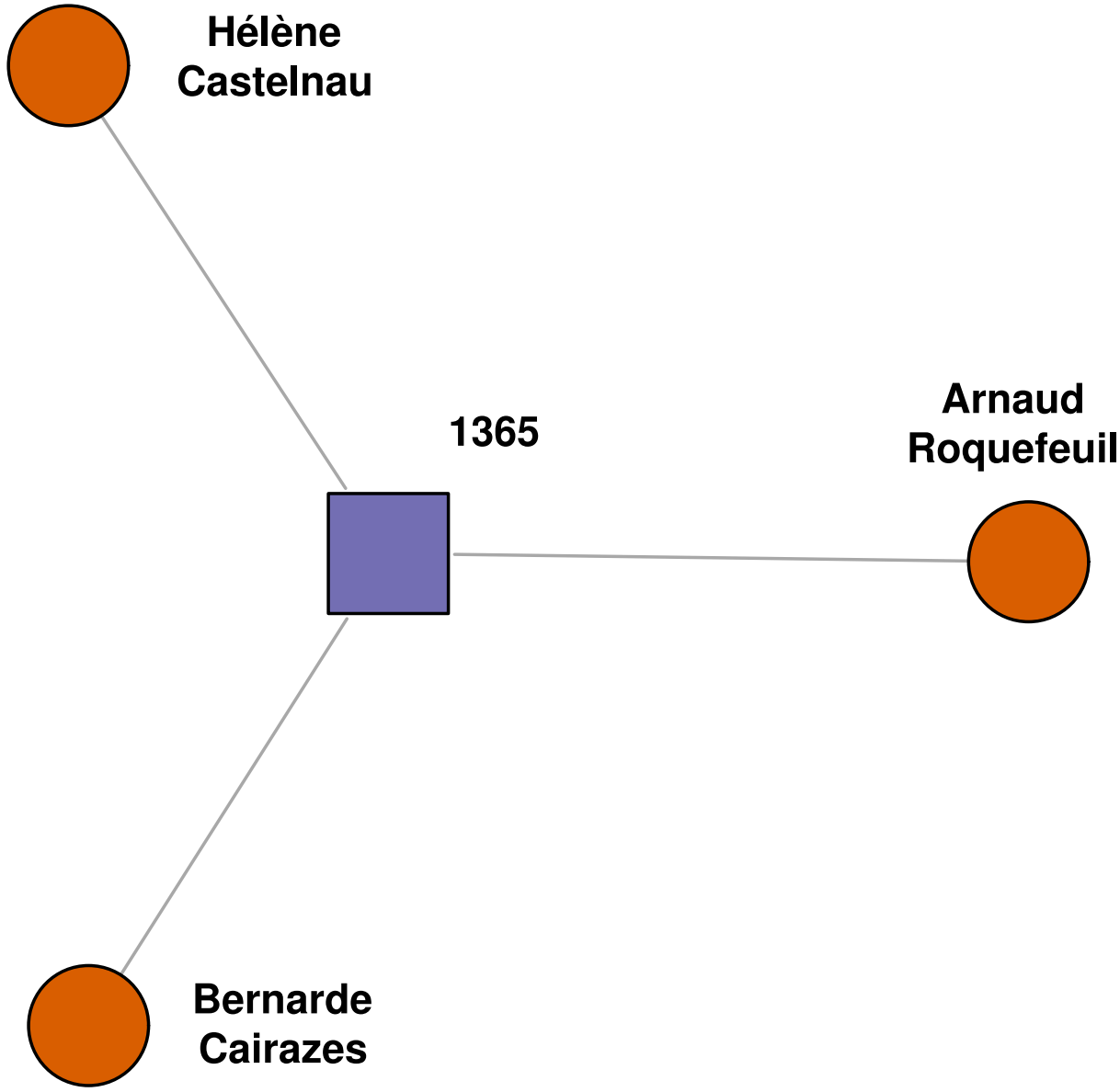
-gestion de la durée dans la représentation des graphes

➔ Quelques propositions en vue de nouvelles analyses de ces données

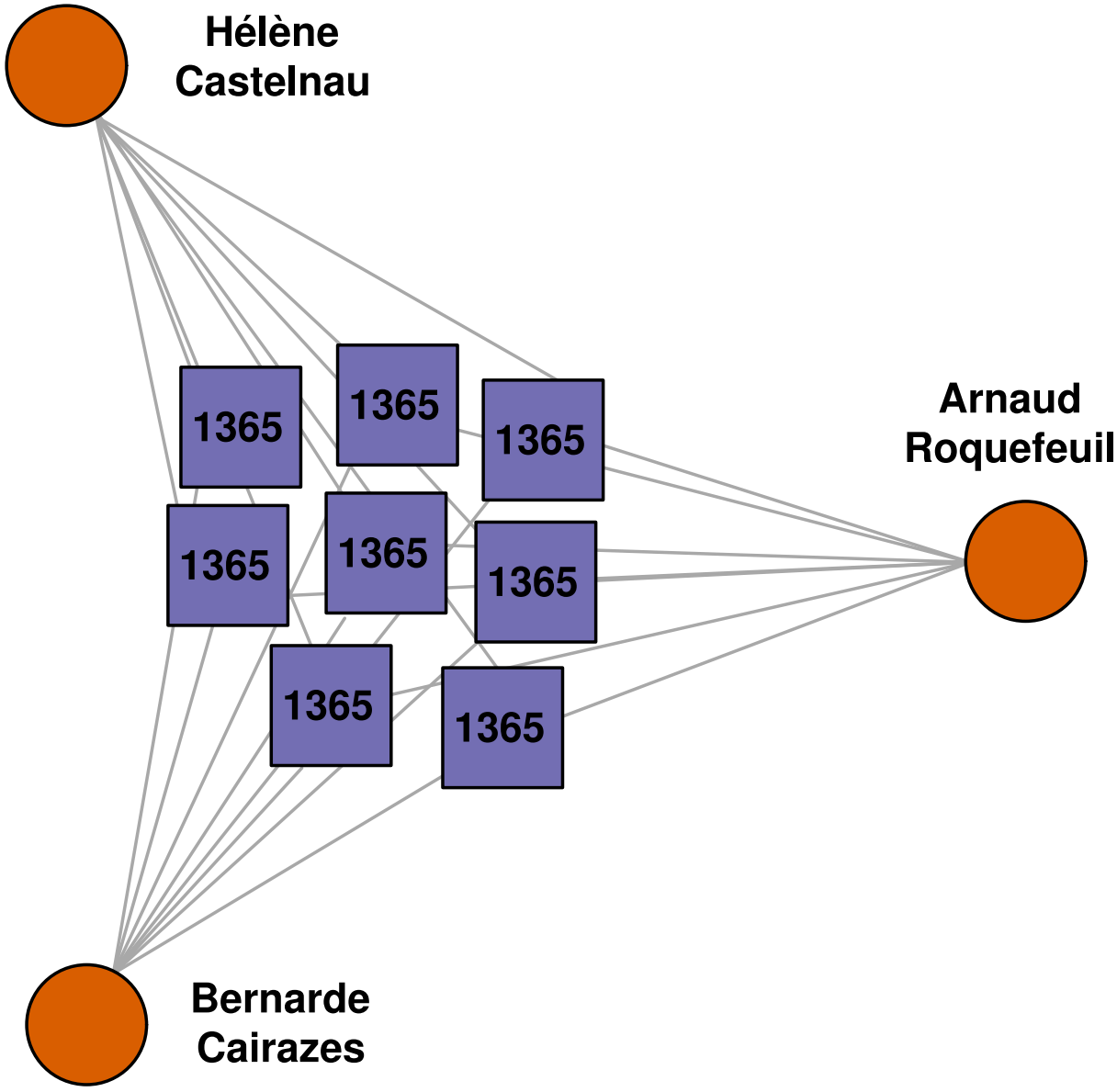
Modèle relationnel

- chaque entité est un sommet d'un graphe : personnes et transactions
- les arêtes du graphe sont les citations : une personne est citée dans une transaction
- le graphe est étiqueté :
  - nom des personnes, « profession »
  - nature de la transaction (bail à fief, investiture, etc.)
  - rente associée
  - nature et localisation du bien
  - date de la transaction
  - nature de la citation (tenant, seigneur)
- proche des modèles des bases de données bibliographiques (type DBLP)

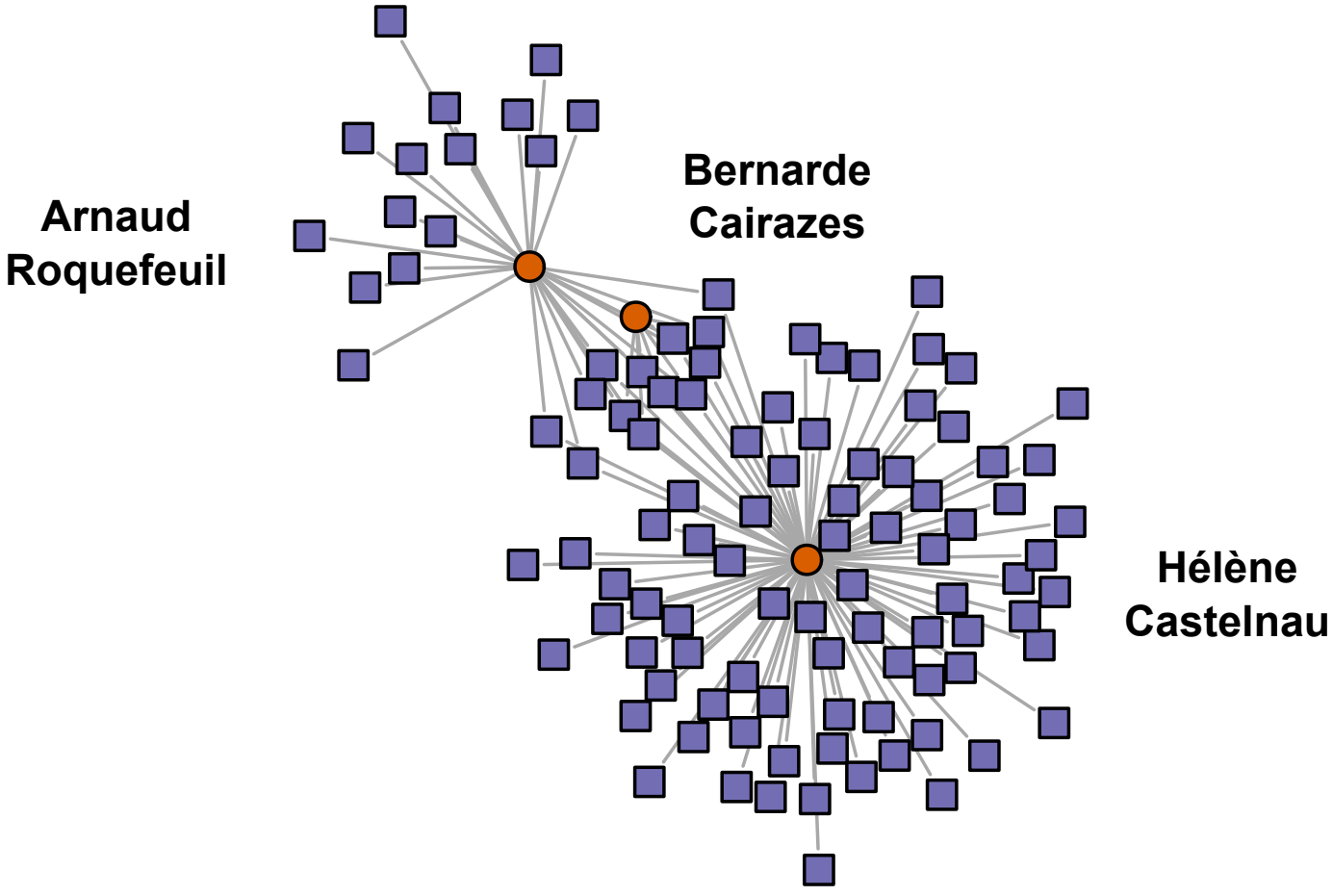
Exemple


Hélène
Castelnau
1365
Arnaud
Roquefeuil
Bernarde
Cairazes

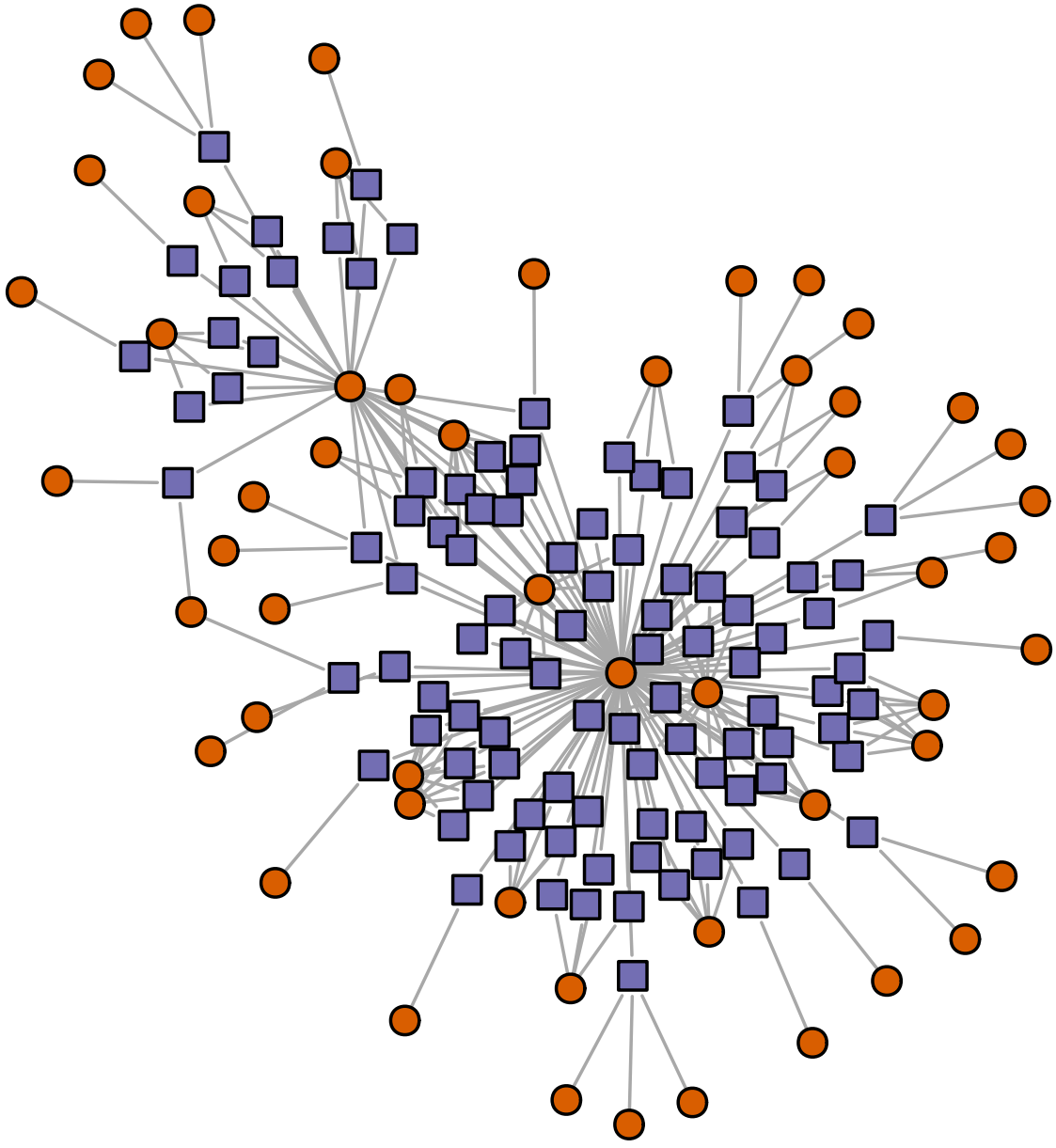
Exemple


Hélène
Castelnau
1365
1365
1365
1365
1365
1365
1365
1365
Arnaud
Roquefeuil
Bernarde
Cairazes

Exemple



Exemple



Modèle

Discussion

Réduire la complexité

- aspects géographiques : localisation des biens par confront
  - « carte » des biens et/ou des lieux ?
  - structure relationnelle des confronts ?
- lien de parenté entre certaines personnes (héritage, mariage, fratrie, etc.)
  - une autre structure relationnelle ?
- les transactions sont regroupées en actes
  - autre type d'entité ?
  - prise en compte des notaires ?

Limiter la perte d'information

- pas de projection du réseau
- pas d'interprétation de l'implication dans une transaction

Difficultés

Taille des données

- 6 487 transactions
- 4 055 noms de seigneurs et de tenanciers
- 10 025 sommets dans une unique composante connexe
- coût algorithmique élevé (parfois inacceptable)

Graphe biparti

- pas de liens entre les transactions et entre les individus
- structure dominante du graphe
- de nombreux modèles ne sont pas adaptés

Difficulté majeure

Les noms

- beaucoup d'homonymes (transmission du nom, par exemple)
- graphie incertaine
- nom de famille pour les nobles mais pas pour les autres personnes
- conséquences :
  - une personne $\neq$ un nom : par ex. 13 personnes différentes sont désignées par « Jean Laperarede »
  - erreurs de transcription :
    - fusion de personnes
    - séparation d'une personne réelle en plusieurs personnes « numériques »
    - mauvaise affectation d'une personne à une transaction
- problème similaire mais moins marqué dans les BD bibliographiques

Quelques éléments d'analyse

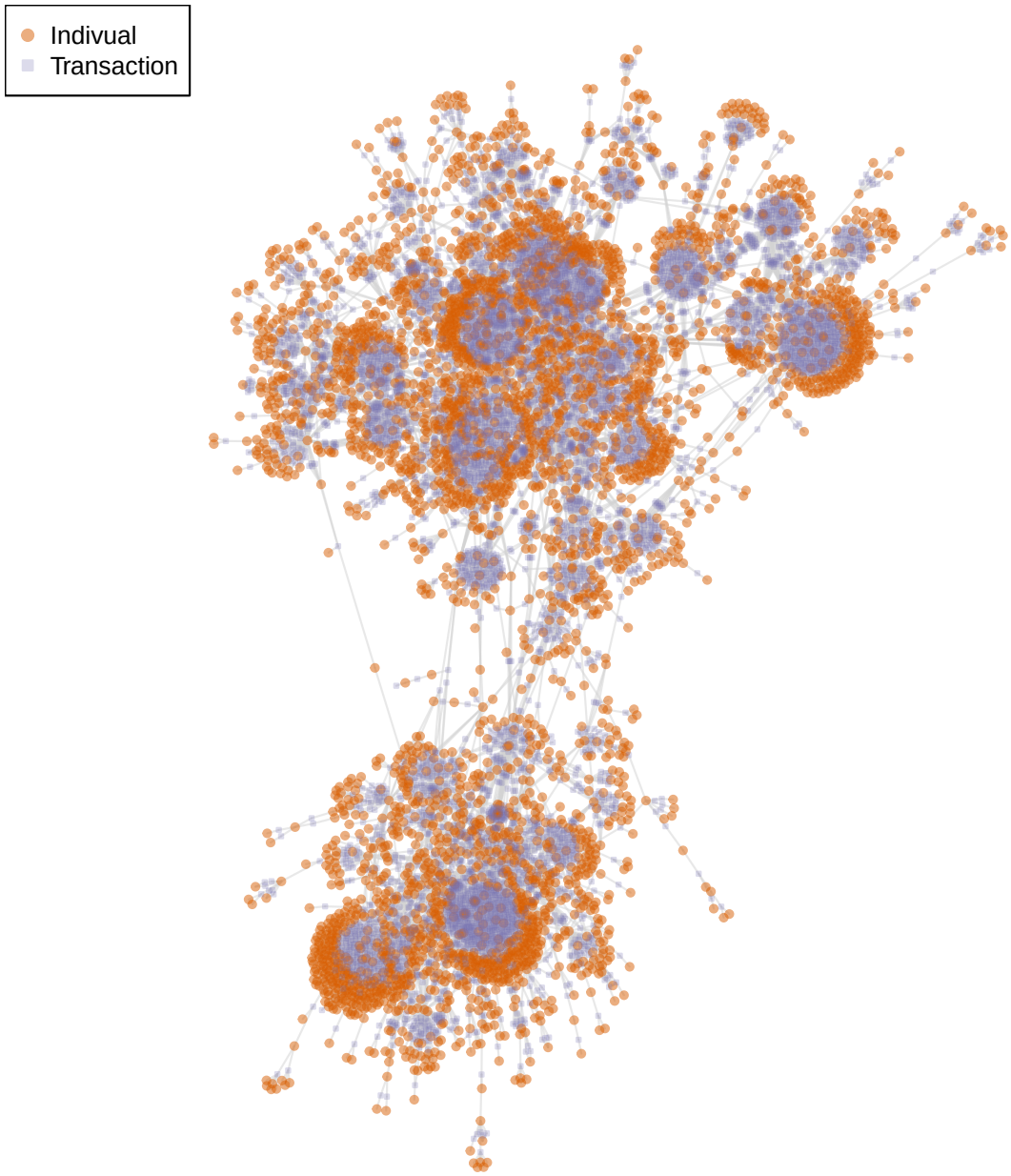
Visualiser le réseau

- 10 000 sommets : limite actuelle du rendu graphique efficace (quelques minutes de calcul)
- visualisation très dense : interaction indispensable

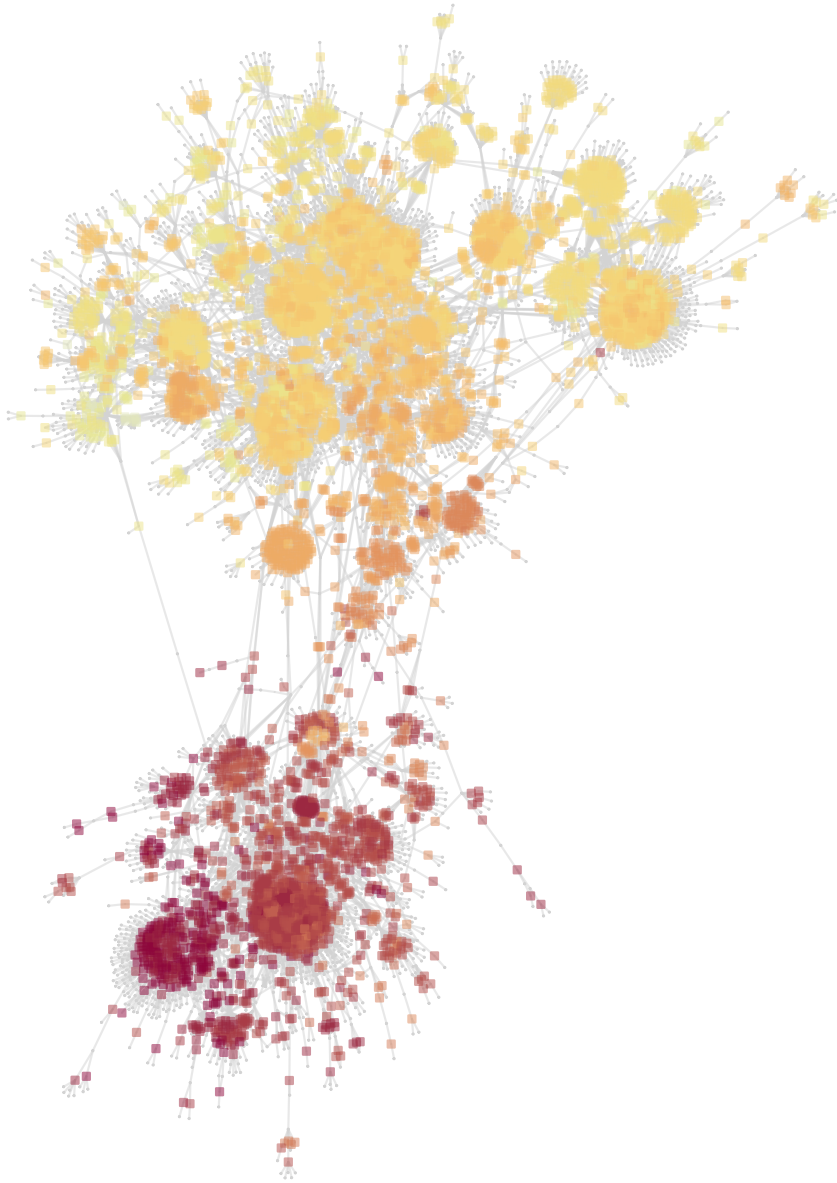
Tester/valider des hypothèses

- approche par simulation
- étudier la fréquence d'apparition d'une structure dans des graphes aléatoires « similaires » au graphe considéré
- exemple :
  - 95.1 % des sommets sont dans une unique composante connexe
  - c'est une couverture plus faible qu'attendu : 98.4 % de couverture sur des réseaux similaires
- enjeux : similarité ? efficacité algorithmique ?

Visualisation

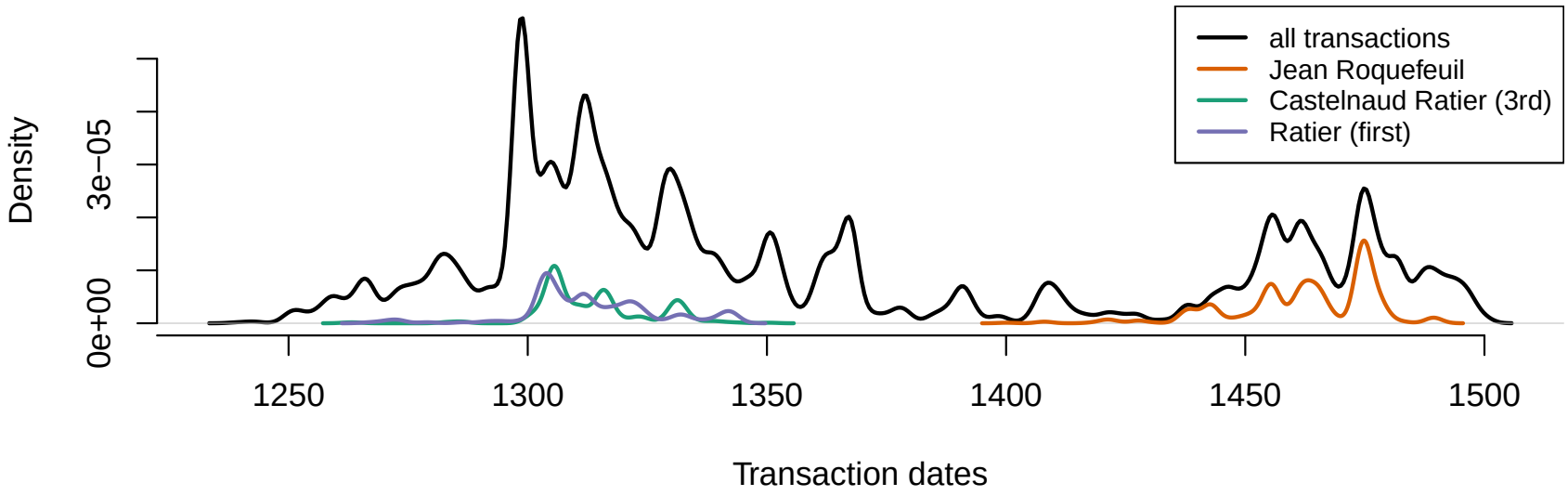


Visualisation



Interprétation historique

Distribution des dates des transactions



- Effets de la peste noire et de la guerre de cents ans
- Deux sous-réseaux :
  - réseau ancien (Seigneur Ratier, Seigneur Castelnau Ratier III, etc.)
  - réseau récent (Jean Roquefeuil, Berenguier Roquefeuil, etc.)

Affiner l'analyse

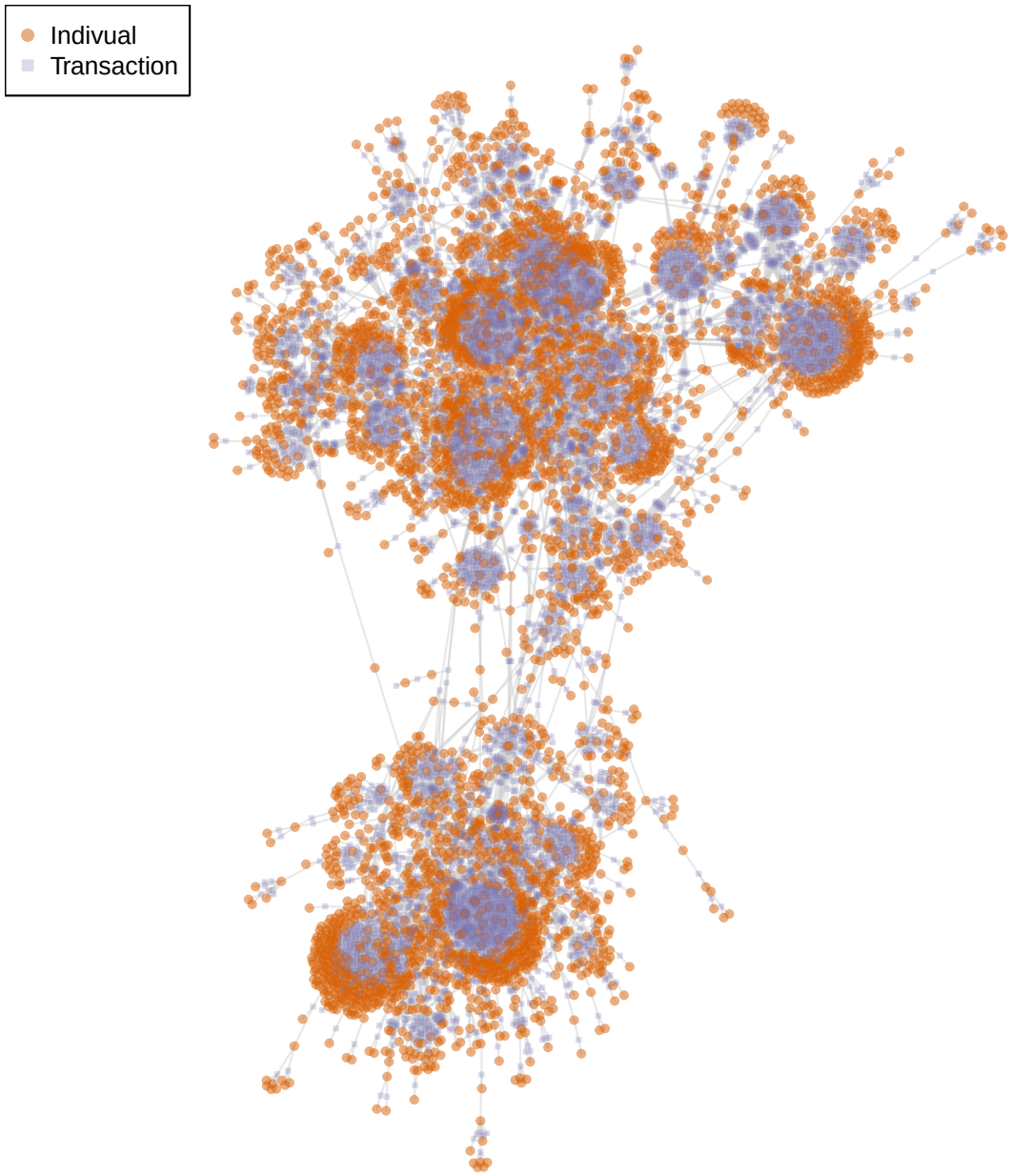
Faciliter l'exploration

- résumer les données
- solution possible : classification des sommets du graphe
- enjeux : objectif de la classification ? structure bipartie ? efficacité algorithmique ?

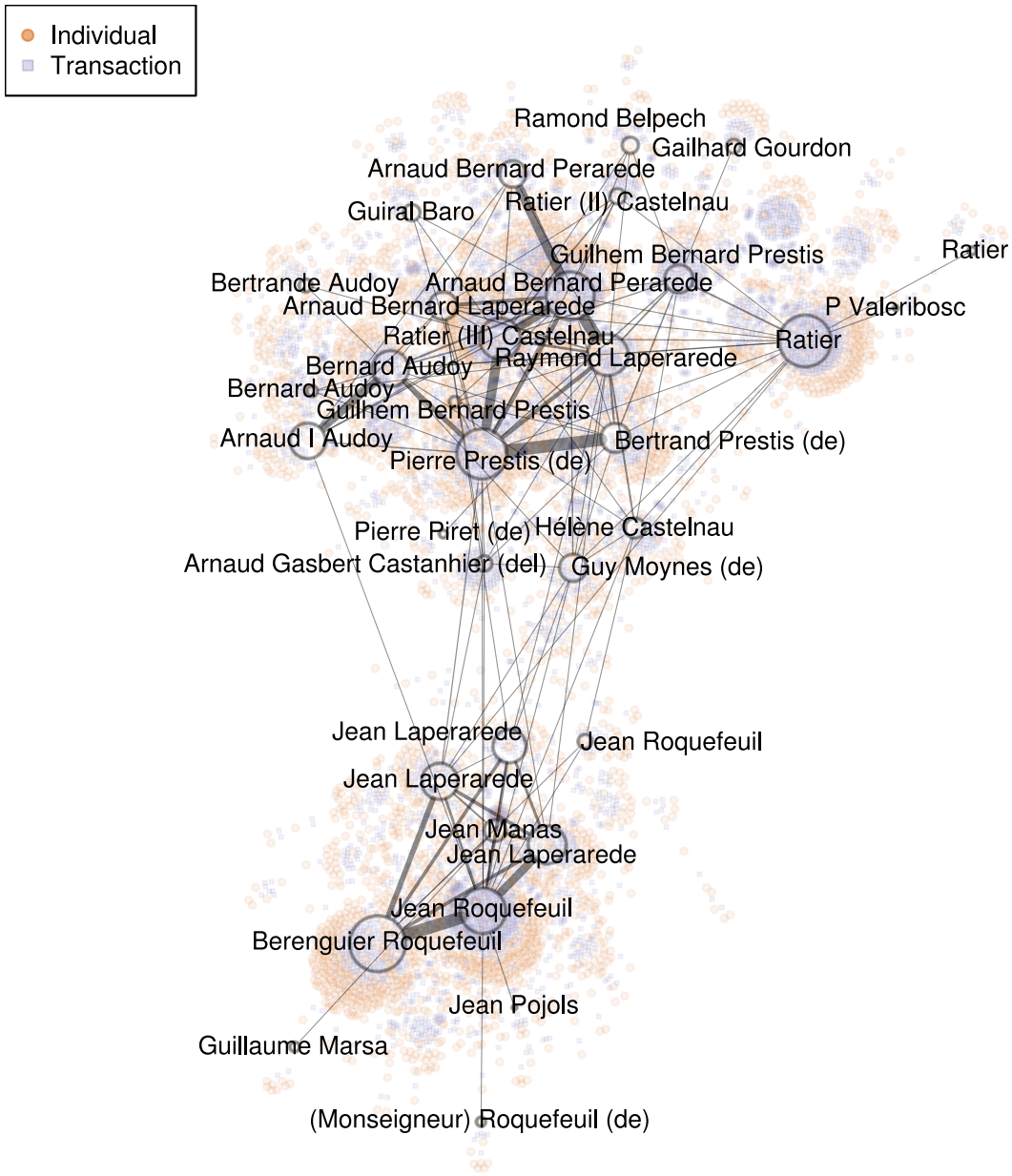
Cohérence des données

- identifier des erreurs de transcription
- détecter des problèmes potentiels
- solution possible : propagation d'informations (par exemple les dates de transaction)
- enjeux : automatisation ? implication de l'analyste ?

Seigneurs



Seigneurs


Individual
Transaction
Ramond Belpech
Gailhard Gourdon
Arnaud Bernard Perarede
Guiral Baro
Ratier (II) Castelnau
Guilhem Bernard Prestis
Ratier
Bertrande Audoy
Arnaud Bernard Perarede
Arnaud Bernard Laperarede
P Valeribosc
Ratier (III) Castelnau
Ratier
Bernard Audoy
Raymond Laperarede
Bernard Audoy
Guilhem Bernard Prestis
Arnaud I Audoy
Bertrand Prestis (de)
Pierre Prestis (de)
Pierre Piret (de)
Hélène Castelnau
Arnaud Gasbert Castanhier (del)
Guy Moynes (de)
Jean Laperarede
Jean Roquefeuil
Jean Laperarede
Jean Manas
Jean Laperarede
Jean Roquefeuil
Berenguier Roquefeuil
Jean Pojols
Guillaume Marsa
(Monseigneur) Roquefeuil (de)

Analyse

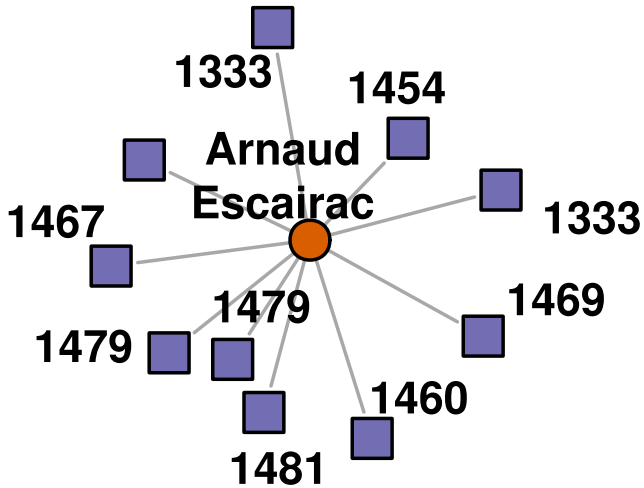
Cohérence

- connexion directe entre Ratier (avant 1350) et Jean Laperarede (après 1350)
- aussi entre Arnaud Audoy et Jean Laperarede
- répétition de noms
- etc.

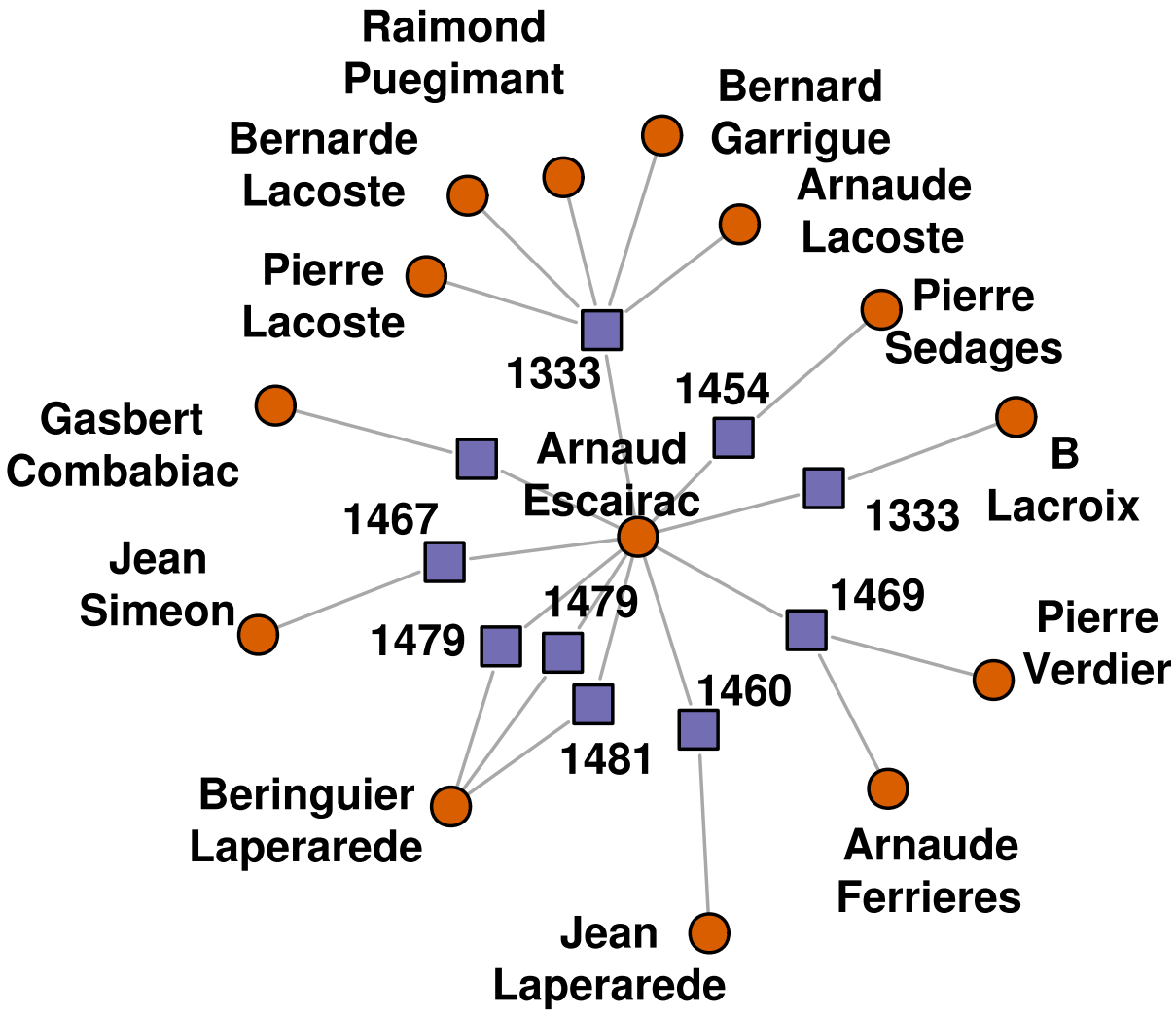
Ratier et Jean Laperarede

- extraction des trois chemins optimaux « directs » entre les deux seigneurs puis des six individus concernés
- propagation des dates de transaction aux individus : durée d'activité suspicieuse pour Arnaud Escairac
- analyse locale

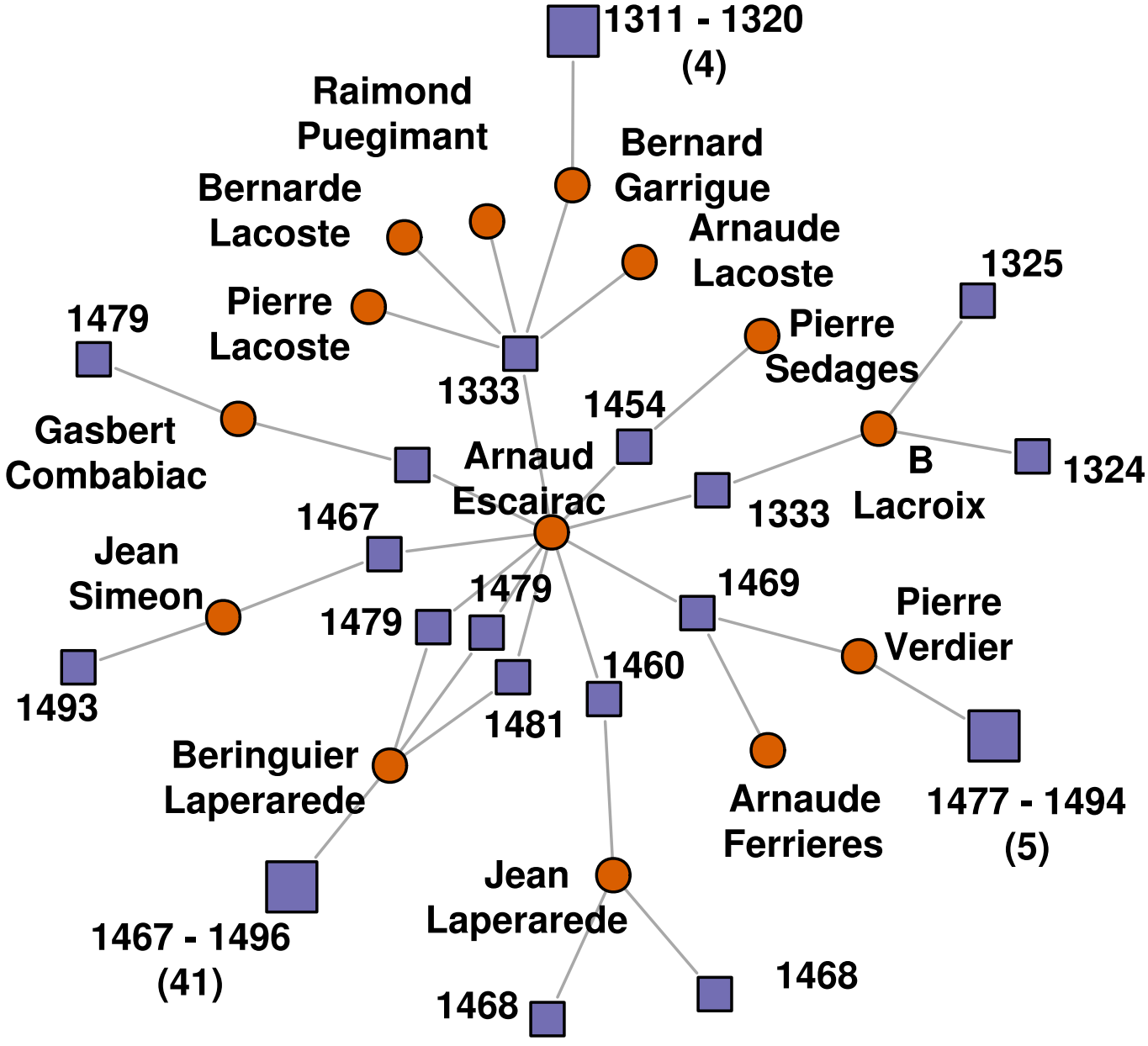
Analyse locale



Analyse locale



Analyse locale



Perspectives

Amélioration des données

- nettoyage de la base de données par traitement semi-automatique des noms
- extraction des informations de parenté

Simplification

- prise en compte de la structure bipartie
- sous-structures fréquentes
- aspects temporels
- visualisation hiérarchique

Validation

- réseaux aléatoires respectant plus la forme du réseau initial
- aspects temporels