Exploration of a Large Database of French Notarial Acts with Social Network Methods*

Fabrice Rossi[1], Nathalie Villa-Vialaneix[1,2] and Florent Hautefeuille[3]

[1] SAMM, Université Paris 1
[2] INRA, UR875 MIA-T
[3] TRACES, Université Toulouse 2

Abstract

This article illustrates how mathematical and statistical tools designed to handle relational data may be useful to help decipher the most important features and defects of a large historical database and to gain knowledge about a corpus made of several thousand documents. Such a relational model is generally enough to address a wide variety of problems, including most databases containing relational tables. In mathematics, it is referred to as a 'network' or a 'graph'. The article's purpose is to emphasize how a relevant relational model of a historical corpus can serve as a theoretical framework which makes available automatic data mining methods designed for graphs. By such methods, for one thing, consistency checking can be performed so as to extract possible transcription errors or interpretation errors during the transcription automatically. Moreover, when the database is so large that a human being is unable to gain much knowledge by even an exhaustive manual exploration, relational data mining can help elucidate the database's main features. First, the macroscopic structure of the relations between entities can be emphasized with the help of network summaries automatically produced by classification methods. A complementary point of view is obtained via local summaries of the relation structure: a set of network-related indicators can be calculated for each entity, singling out, for instance, highly connected entities. Finally, visualisation methods dedicated to graphs can be used to give the user an intuitive understanding of the database. Additional information can be superimposed on such network visualisations, making it possible intuitively to link the relations between entities using attributes that describe each entity. This overall approach is here illustrated with a huge corpus of medieval notarial acts, containing several thousand transactions and involving a comparable number of persons.

Keywords: relational data; network analysis; transcription error detection; notarial acts; data mining in graphs; clustering in graphs

*This work was partially supported by ANR project 'Graph-Comp', ref ANR-05-BLAN-0229. The authors would like to thank Jonathan Jarret and Nicholas Szczesniak for their thorough proofreading of the article as well as the two anonymous reviewers for their helpful comments and suggestions.

1 Introduction

The main objective of this article is to illustrate how mathematical and statistical tools designed to handle relational data may be useful to help decipher the most important features and defects of a large historical database and to gain knowledge about a corpus made of several thousand documents. In this article, 'relational data' means data where the entities under study are described not only in numerical terms or by reason of their intrinsic qualities, but also by the way they are connected to each other. For instance, in the notarial acts considered in this paper, the entities under study are the persons actively involved in the acts. Several facts can be extracted from the acts about those persons, such as their names, their occupations, their ages, and so forth. Additionally, though, two persons can be said to be 'related' if they take part, in whatever way it may be, in the same act. In mathematics, such a relational model is referred to as a 'network' or a 'graph'. Hence, in this article, the word 'network' describes not only what is commonly called a 'social network' but more generally any kind of relational data. Similarly, the term 'graph' should not be here understood as signifying a graphical representation, but only as the mathematical object that models this relational data. Such a relational model is general enough to address a wide variety of problems. Its use is understandably common in a social network framework but its application is certainly not restricted to this field. On the contrary, it is suited to most databases containing relational tables. Thus, mathematical tools associated with this model and mostly developed in a social network framework can be used to extract information from other such databases as well, for instance from citations databases (for articles or patents). Examples and references of the use of graphs as models of various real-life interactions, ranging from collaboration networks to epidemic propagations can be found in (Dorogovtsev and Mendes, 2006, p. 31-83). In historical research, networks are used more and more frequently (see Rosé (2011) or the numerous references on the research platform `https://oeaw.academia.edu/TopographiesofEntanglements` for examples of the use of networks in History or Bertrand et al. (2011), Lemercier (2012), for a general discussion on this topic) but most of these studies use the network as a convenient and intuitive way to represent a set of interactions that are almost exclusively social interactions between people or countries. They remain generally unaware of the available mathematical tools that can help gain a clearer understanding, once the model is built. Except for some individual characteristics of the entities in the network (e.g., the degree or the betweenness, see section 4 or Rosé (2011)), these tools are rarely used to understand the network's main features and almost never combined to check the consistency of the data.

This article's purpose is to emphasize how a relevant relational model of a historical corpus can serve as a theoretical framework which makes available automatic data mining methods designed for graphs. By such methods, for one thing, consistency checking can be performed so as to extract possible transcription errors or interpretation errors during the transcription automatically. We differentiate between transcription errors, due to faults when the original text is copied into the database, and interpretation errors, due to erroneous interpretation of the text. For example, a very challenging issue for historians and digital medievalists trying to create prosopographical databases is that until the early

modern period, 90% of the population was without surnames. Those who had surnames and Christian names often bore the same names as their ancestors, leading to a large number of namesakes. Keats-Rohan (2007) includes several chapters discussing this issue and providing methodological hints for addressing it (see, in particular, pages 95-230 in the section "Planning a prosopography: possibilities and problems"). The present paper does not intend to compete with the present strategies used to differentiate persons with identical names, but it aims to illustrate how an analysis based on a network model can complement these strategies, and uncover possible erroneous interpretations made by the historians who transcribed the documents into the database. Such mistakes have led to mergers of two distinct individuals with identical names or, on the contrary to the splitting of one individual into two distinct entries in the database. Moreover, when the database is so large that a human being is unable to gain much knowledge by even an exhaustive manual exploration, relational data mining can help elucidate the database's main features. First, the macroscopic structure of the relations between entities can be emphasized with the help of network summaries automatically produced by classification methods, as will be explained below. A complementary point of view is obtained via local summaries of the relation structure: a set of network-related indicators can be calculated for each entity, singling out, for instance, highly connected entities. Finally, visualisation methods dedicated to graphs can be used to give the user an intuitive understanding of the database. Additional information can be superimposed on such network visualisations, making it possible intuitively to link the relations between entities using attributes that describe each entity.

This overall approach is here illustrated with a huge corpus of medieval notarial acts, containing several thousand transactions and involving a comparable number of persons. The whole corpus has been recorded in a database and contains largely similar sorts of transaction from a closely restricted geographical area, giving considerable homogeneity of data. An induced graph has been derived from the corpus, relating the persons involved in the transactions to the transactions themselves. This graph contains more than ten thousand entities (both transactions and persons). The paper is organized as follows: the section 2 describes the corpus, its associated database and the relational model derived. The section 3 focuses on the global analysis of the network, using statistical indicators, visualisation techniques, and clustering methods. The section 4 illustrates the use of local numerical indicators for discovering important persons in the network. The section 5 shows how information propagation within networks leads to semi-automatic consistency checking when coupled with local visualisation. We end with a conclusion summarizing the benefits of the methodology.

2 Data description and modelling

The corpus just introduced is physically preserved at the Archives départementales du Lot (Cahors, France) and is available for public consultation Miquel and Willy L. (2011). The corpus is divided into four registers, shelfmarks AD 46 48 J 3, AD 46 48 J 4, AD 46 48 J 5 and AD 46 48 J 6. The corpus is the work of a feudist who was hired for twenty years in the eighteenth century to collect all the notarial acts he could find that mentioned the successive lords of

the seigneurie Castelnau Montratier. This work was designed to help the new owner of the seigneurie, Jean-Léon de Bonal (a former bourgeois, ennobled), to claim his rights over lands and collect rents from his new properties, and provides us with a substantial corpus of documents that have otherwise been completely lost, since the originals do not survive.

The documents are notarial acts, each containing one or more transactions. The corpus is homogeneous from several points of view. The transactions are all related to the *seigneurie* Castelnau Montratier, near the present-day village of the same name (le Lot, south-west France). About forty parishes are included in the *seigneurie*, which covered a total area of approximately 300 km^2. All the acts are of similar types: they are all notarial acts describing agreements of different sorts made within the *seigneurie* (purchase, sale, donation, tenancy, manumission, dowry...). The majority of these acts concern land. The original transactions took place between 1238 and 1768, with their abundance shifting over the period. The whole corpus can be seen as a very representative, if not exhaustive, sample of the land charters written in this *seigneurie* during that period.

The transactions are recorded in a large database freely available online Hautefeuille (2009). More precisely, more than 75% of the whole corpus has already been digitized, the precise amount varying from register to register: priority has been given to an homogeneous geographical sample and to transactions dating from before 1500. As an example, the act partially reproduced in Figure 1 contains a transaction transcribed below:

> *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.*
> **[. . .] (seven other transactions for two gardens, a meadow and four plots of land)**
> *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.*[1]

The transaction contains various data, such as the act reference and page, *AD4648 J6 page 37, acte 26* (in the margin), the transaction date, *1365, le*

[1] '1365, the Wednesday before Pentecost. Enfeoffment by my lord Arnaud of Roquefeuil and Lady Hélène of Castelnaux his wife in favour of Bernarde of Cayrazes, daughter of the late Arnaud of the parish of St-Jean de Cornus, of a house at la Graulière, in the said parish of Cornus, touching on one side the land of Jean of Cayrazes and on two sides the public roads of the said place of la Graulière. . . . subject to the render of two Cahorcin sous in recognition of the change of lord or of feudatory and of 3 emines of barley - an emine is worth a half-sétier and the sétier 4 quarts - and one chicken on Lady Day in September. Jean de Combelcau, notary and Commissary of Authority of my lord the official of Cahors.' The picture is reproduced with the kind permission of the Archives départementales du Lot, copyright Florent Hautefeuille, 2005.

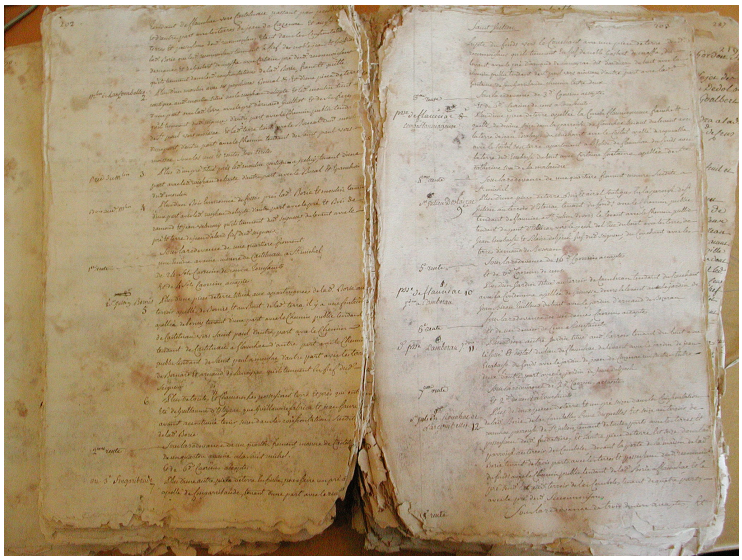

Figure 1: Example of an act (partially reproduced) including several transactions (tenant farming, recorded as transaction ID 142 in the database)

mercredi avant la Pentecôte (also in the margin), the lords directly involved in the transaction, *Arnaud de Roquefeuil* and *Dame Hélène de Castelnau* (his spouse), the tenant directly involved in the transaction, *Bernarde Cayrazes*, the location of the land concerned, *La Graulière, paroisse de Cornus*, located by the place's name and its parish, the neighbour of the land concerned, *Jean Cayrazes* and the notary who wrote the transaction out, *Jean de Combelcau*. (Other data are recorded about this transaction but for the sake of simplicity, only the information used in the remainder of the paper is mentioned here.)

A relational model is next derived from the database: the 'vertices' of the graph, modelling the entities under study, are both the transactions and the individuals directly involved in those transactions. The relations between entities are modelled by 'edges' that connect some pairs of vertices. Here two vertices are connected when the individual represented by one vertex is directly involved in the transaction represented by the other vertex. Hence, in this graph, an edge only connects one individual and one transaction. This kind of graph is said to be 'bipartite'. Definitions of graph-related terms can be found in introductory books on graph theory such as Voloshin (2009) or in less formal terms in books such as Scott (2000). Figure 5 gives an illustration of one tiny part of this relational model. In these figures, a individual named Guiral Combe is involved in five transactions (transaction dates are given on the left sub-figure while the right one displays the parishes to which the place concerned by each transaction belongs). Two of these transactions were made with a Jean Laperarede, one was made with a Guilhem Bernard Prestis, another one with three other individuals named Pierre, Guillem and Raymond Laperarede and the last one is a transaction where only Guiral Combe is mentioned.

Our model is restricted to transactions from before the year 1500 to avoid distortion to the graph due to the low proportion of transactions after that date so far digitized. The whole final graph contains 10,542 vertices (6,487

transactions involving at least one individual and 4,055 individuals involved in at least one transaction as an active participant, either tenant or lord).

3 Global network analysis

Once a relational model has been defined, the user usually wants to use it for answering questions that he may have regarding the content of the corpus in order to gain specialized (i.e., historical) knowledge. Commonly, a first stage would be to identify the key social actors among the 4,055 individuals, and to obtain an overall description of the relations between these actors. For such aims, a network model is better suited than traditional individual analyses of the people involved in the transactions since it explicitly provides an explicit global overview of the relations between individuals. However, visualizing these relations, even after the model has been set, is not a straightforward process in a network that has several thousand nodes. Indeed, a relational model does not come with a "natural" visualisation and several techniques can be used to display a network. Some open-source graph exploration software, such as Gephi[2], implement some of these techniques and provide interactive graph exploration tools. The choice of one or another visualisation technique can eventually induce interpretation bias,but this problem can be limited by combining the visualisation with other statistical analyses, which serve as validation tools for highlighting the most important facts (and relations) in the network. This section describes how such an analysis can be conducted and which kind of results can be obtained.

Before visualisation, a global analysis generally starts with a study of the connectivity of the graph, answering the following question: 'can any single vertex of the graph be reached from any other vertex following edges?' When the answer is positive, the graph is said to be connected. Otherwise, the disconnected graph is made of connected components, which are maximally connected sub-graphs.[3] As those components are disconnected, they can be analysed independently.

The network under study is not connected thus, but it contains a very large connected component that comprises 3,755 individuals and 6,270 transactions, that is, 95.1% of the vertices of the full graph. Perhaps surprisingly, this coverage is significantly smaller than expected: using computer-based simulations, as in Kannan et al. (1999), one can show that, on average, the largest connected component occupies 98.4% of the vertices of graphs with a similar macroscopic structure (i.e., bipartite graphs with a specific degree distribution). In simpler terms, this means that our graph of notarial acts is unexpectedly poorly connected: rather than having some kind of overall uniformity, it contains parts with connectivity patterns denser or sparser than expected. In addition to the dominating component, the notarial graph contains 107 very small components (with fewer than 11 persons in each). These small components will not be analysed any further in this paper but could have been visualized separately in the

[2]The software is available from Gephi - Makes Graphs Handy http://www.gephi.org, see Bastian et al. (2009).

[3]Connected components are obtained by following the links: here one starts from a random person and moves to transactions in which this person plays a role, then to other persons involved in those transactions, and so on until the whole graph has been travelled or this has proved impossible.

same manner.

Figure 2 provides a representation of the largest component of the notarial graph. This figure is generated by a two-step Fruchterman and Reingold-like algorithm Fruchterman and Reingold (1991), similar to the one proposed in Tunkelang's PhD thesis Tunkelang (1999). Even if reading the fine details of this static picture is difficult, broad structures are very obvious. In particular, Figure 2 shows that the graph contains two loosely connected parts (the upper and lower parts of the figure), which themselves have clear substructures. The visualisation seems to confirm the poor connectivity structure discovered above: substructures are densely connected internally and weakly connected to other substructures.

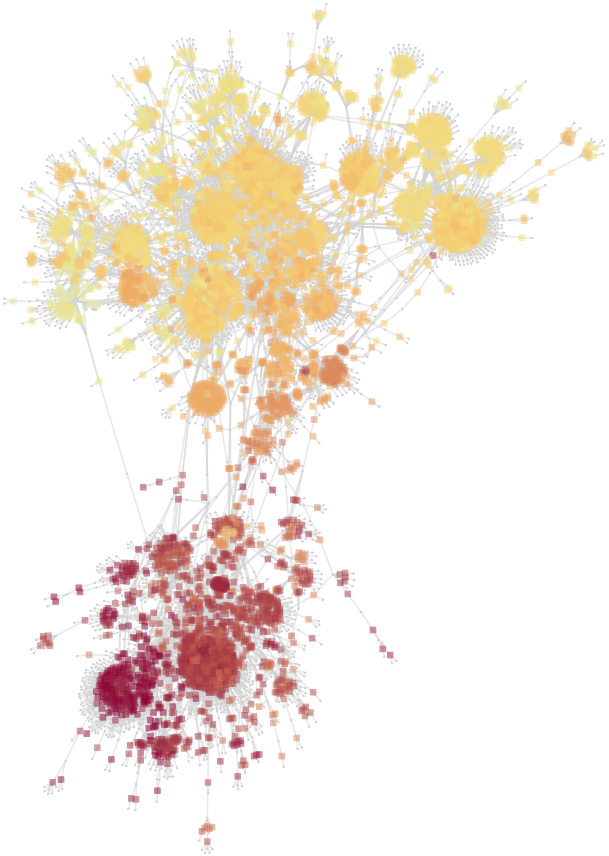

Figure 2: Largest connected component of the bipartite graph visualised by means of a force-directed placement algorithm. Transactions are displayed by large squares and individuals by small circles. Colours encode the transaction dates (red is for more recent dates and yellow for older ones).

In fact, as shown by the colour-based representation of the transaction dates, there is a very good agreement between the graph visualisation (which is not computed using those dates, although it shows them) and the temporal aspects

of the database: close transaction vertices on the Figure are temporally close and vice versa. Thus, the poor global connectivity can be easily explained by the transaction date distribution (see Figure 3): notarial activity is very scant around 1400, as a consequence of the Black Death, and during the Hundred Years War period. Therefore, the older period of the notarial network (upper part of Figure 2) and the more recent period (lower part of Figure 2) are only poorly connected by the intermittent transactions that took place in the period around 1400.

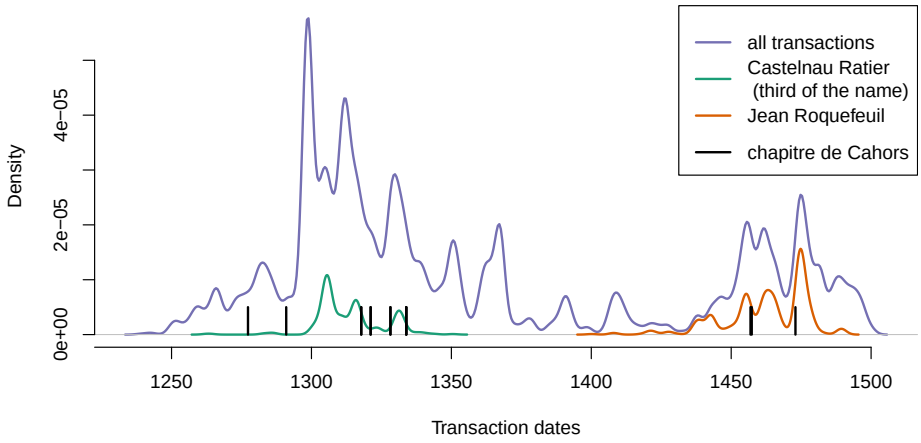

Figure 3: Transaction dates distribution with the three main lords.

The combination of traditional techniques (representation of transaction dates) with graph-related methods (connectivity and visualisation) leads here to a deeper insight into the archives than could be gained from each approach used independently. Indeed, as Figure 2 appears reliable, because of the strong correlation between transaction dates and transaction vertex positions, dense substructures are very probably meaningful, especially as they demonstrate the poor connectivity which has been established by the simulation method mentioned above. Without the confirmations by dates and the simulation, the reliability of Figure 2 might be questionable. Conversely, without some attempt at graphing, it would be hard to identify periods of dense notarial activity from the transaction date distribution alone.

Nonetheless, Figure 2 remains very complex. It can be explored with interactive software that supports zooming and panning, but the user is likely to be overwhelmed by the size of the graph. A common approach for managing the complexity of large databases consists in using clustering methods. In the graph context, clustering aims at partitioning the vertices into groups that are densely connected, such that vertices belonging to two different groups are comparatively more poorly linked. Thorough overviews of vertex clustering methods are given in Fortunato (2010) and Schaeffer (2007).

Graph vertex clustering is used here to enhance Figure 2. Clustering bipartite graphs in a meaningful way remains an open issue. When the two types of vertices have comparable connectivity properties in the network, one can build independent but consistent clusters of each type of vertices Barber (2007). In the present case, this would lead to clusters of transactions and to clusters of

persons. However, we will show that transactions have very different connectivity properties than persons, a fact that would introduce strong distortions in this approach. We use therefore a simpler solution in which a recent clustering technique is applied only to persons. The technique is described in details in Rossi and Villa-Vialaneix (2011). More precisely, a projected graph is constructed. It contains only the vertices associated with persons; rather than associating them directly to transactions, this graph contains an edge for each pair of persons that appear in the same transaction. The final clustering contains 34 clusters whose size varies from 2 to 400 persons. The mean number of persons in a cluster is about 110. Then, a central individual is identified in each cluster: this is the person who appears in the greatest number of transactions.

Figure 4 shows how the clusters and the central persons can be used to improve Figure 2. The 34 central persons are marked on the graph and identified by their name. Each cluster is materialised using a circle that encloses an area proportional to the number of vertices included in the cluster; the circle is centred on the central individual. The connectivity structure of the graph is summarized using edges between clusters: the width of each such link is proportional to the number of transactions between members of the clusters that it links.

This summary provides a much clearer global perspective on the notarial graph than the original picture. For instance, it confirms the specific connectivity pattern of the network, since edges appear concentrated between some clusters rather than evenly spread. It points the viewer toward transactions and persons that connect the older period of the graph to the more recent one, for instance to Hélène Castelnau, Guy de Moynes and Arnaud Gasbert del Castanhier.

While the summary does not replace a detailed local exploration of the network, it provides a simplified global map that can be referred to when zooming on details. It guides also the exploration by emphasizing possible issues in the database. There are, for instance, suspicious multiple occurrences of identical names (see, e.g. the two 'Ratier' clusters on the top right of the figure). The direct connection between the large 'Ratier' cluster on the top right and one of the clusters whose central person is named 'Jean Laperarede' on the bottom left is also surprising as those clusters belong to different time periods.

This particular issue can be analysed further with graph techniques. It should be first noted that clusters used in Figure 4 are guaranteed to be internally connected as a result of the clustering method used. The display of a connection between two clusters therefore corresponds to the existence of a direct 'path' between every vertex of the first cluster and every vertex of the second one.[4] As the graph is connected, the existence of general paths between any pair of vertices is guaranteed: the surprising aspect lies here in the fact that the connection is direct from an old cluster to a more recent one. We, however, are not interested in any path but the most direct, the 'shortest' paths between Ratier and Jean Laperarede.[5] Among these, only three paths go directly from the Ratier cluster to the Jean Laperarede one, and all link Ratier to Jean Laperarede via Bernard Garrigue, Arnaud Escairac, Berenguier

[4] In a graph, a 'path' is a sequence of vertices such that from each of its vertices there is an edge to the next vertex in the sequence.

[5] The shortest path between two vertices is the path (or sometimes the paths) between those vertices featuring the smallest total number of edges.

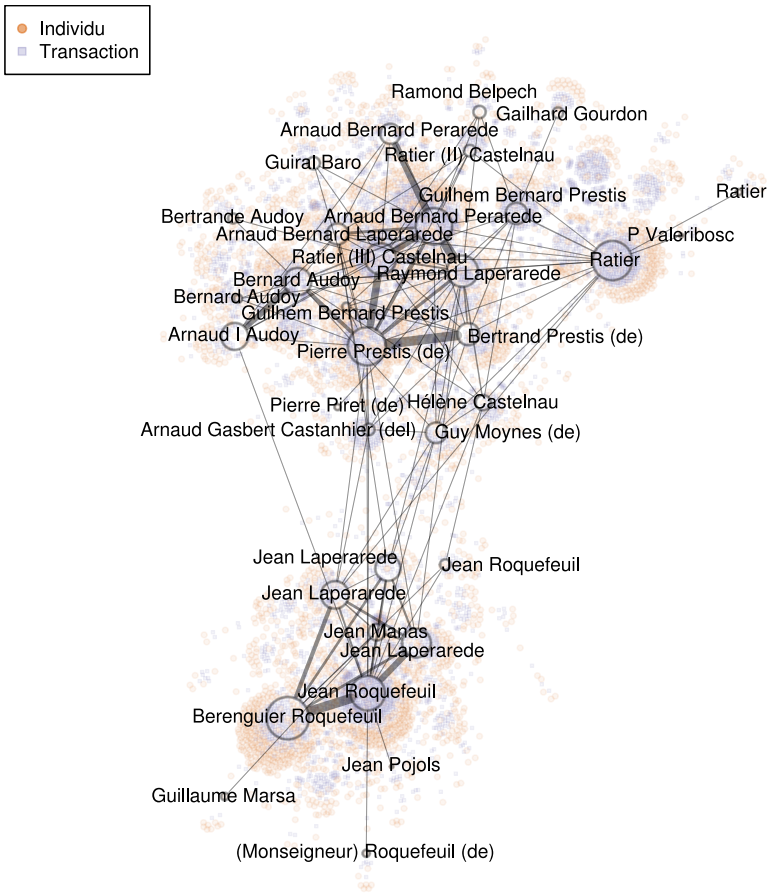

Figure 4: Largest connected component of the bipartite graph with a simple representation of the maximal modularity clustering result. Each circle corresponds to a cluster and has an area proportional to the cluster size. The circle centres are positioned around the vertex with the largest degree (the 'leader' of the cluster).

Laperarede and a fourth person that differs in each path. This analysis turns a suspicious visual detail into a list of persons that should be studied in detail to verify the consistency of the database. We will postpone this verification to a later section 5, after having introduced other automated means of singling out important persons in the graph.

4 Local network analysis

The global analysis conducted in the previous section provides a general view of the database together with some hints about possible substructures, transcription errors, or interpretation errors during the transcription. These aspects must be confirmed by more finely grained techniques. In particular, one of the main findings of Section 3 is the specific connectivity structure of the notarial

graph, which is itself a consequence of the distribution of transaction dates. This section addresses the analysis of the connectivity by focusing on numerical characteristics obtained at the node level, such as the 'degree' or the 'betweenness'. More precisely, we demonstrate that node characteristics are related to different aspects of the 'importance' of an individual. Some of these aspects could have been obtained directly by studying the corpus in a standard way (e.g., looking at the number of transactions in which each individual is involved) but others are very specific to the network model and are based on a global examination of the relationships (e.g., using the previous clustering or defining a centrality measure for the individuals). Combining the two approaches yields an identification of key actors, singular patterns, transcription errors or possible interpretation errors in the transcription that could not have been found as easily with a standard approach, which would not have taken the global network structure into account.

The 'degree' of a vertex is the number of edges pertaining to the vertex. In the notarial graph, this corresponds to the number of transactions in which an individual is involved, for person vertices, and to the number of persons involved in a transaction for transactions vertices. Hence, the degree is a measure of the popularity of a vertex in the network. In the notarial graph, the degrees of transaction vertices are very different from the degree of the person vertices, as is to be expected: transactions are inherently limited to a few persons, up to twelve in the database. On the contrary, degrees of persons exhibit the classical power law behaviour observed in many real world networks: most of the person vertices here have a very low degree (1,815 individuals appear in only one transaction each) but a handful of vertices have very high degree Barabási and Albert (1999). (The extreme case is Jean Roquefeuil, who appears in 551 transactions.) As might be expected, the individuals with high degree were all nobles and appear as seigneurs in most of the transactions in which they are involved.

With the notarial acts, a network approach is not needed to study the degree of the persons; it is a natural classical measure of the activity of the individuals under study. While interesting facts can be observed using this quantity (for example, the fact that two women, Lombarde Laperarede and Hélène Castelnau, appear in the top twenty-three individuals), we are more interested here in findings that cannot be obtained without using the graph model. This is the case with discrepancies between the list of top-degree individuals and the list of 'leaders' obtained in the previous Section. Those two lists agree to some extent as they share twenty-one persons (of thirty-four), but they have noticeable differences. For instance, Raimond Perarede, who is the sixth person in degree order and who appears in 234 transactions, is not considered as a leader by the clustering analysis: he has been included in the same group of persons as Arnaud Bernard Perarede, who appears in 304 transactions. This is explained by the large number of transactions (fifty-one) that involve both seigneurs. This pattern repeats itself twice: for Bernard Audoy and Jacmes Audoy (116 transactions in common) and for Raymond Laperarede and Gausbert Lauriac (52 transactions).

These discrepancies reveal interesting transaction patterns and family relationships in the corpus. Bernard and Jacmes Audoy were indeed brothers and had inherited common lands and rights from their father (also named Bernard Audoy): they then made common contracts with tenants. (The Audoy family

is well represented in the upper left part of the graph on Figure 4.) The two other pairs, however, Arnaud Bernard Perarede and Raimond Perarede on the one hand and Raymond Laperarede and Gausbert Lauriac on the other, are explained by different patterns in the corpus: these collaborators made a few acts (two or three) that comprised a large number of transactions (about thirty transactions together in at least one case). These acts are large exchanges (between Arnaud Bernard Perarede and Raimond Perarede) or sales (between Raymond Laperarede and Gausbert Lauriac) that may correspond to an important local change in the social organization.

Additionally, as some high-degree persons do not appear in the leader list, lower-degree persons get somehow promoted: thirteen clusters have a leader who appears in less than thirty-four transactions whereas the thirty-fourth degree in decreasing order is fifty-three. These leaders are associated with small clusters and this points to possible interpretation errors in the transcription. For instance, there are two 'Guilhem Bernard Prestis' associated with two distinct clusters. One of the individuals has a high degree (204) while the other one has a small degree (12) and is the leader of a small cluster. It is probable that those two vertices are in fact the same person. Another standard numerical characteristic in social network analysis is the vertex 'betweenness'. This is the number of the total of shortest paths between all pairs of vertices that pass through the vertex in question. Betweenness is then a centrality measure: vertices with a large betweenness are likely to disconnect the network if removed. Contrarily to the degree, the betweenness is a non local measure that cannot be defined outside of a graph structure.

Once again, the top betweenness individuals list is very similar to the top-degree individuals list: among thirty-four individuals, the two lists have twenty-four persons in common. Individuals with a large betweenness who are not in the top degree list should be analysed with a special focus. For some of them, the centrality is easily understood. *Chapitre de Cahors* (i.e., the cathedral chapter of Cahors) or *Église de Flaugnac* (church of Flaugnac) have a large betweenness because they are not mortal persons; they were corporations that got involved in transactions over many different periods. Thus, they link the older period of the archives to the newer one. For other persons, a more subtle analysis is needed, which uses the network structure to a larger extent than the previous analyses.

5 Information propagation

Transactions in the notarial acts graph come with numerous associated characteristics, in particular the date of the transaction and the parish associated to it. This translates automatically into temporal and geographical ranges of activity for persons, by associating with persons the information from the transactions in which they took part. A standard approach to take the benefit of these informations would be to use them at the person level: identifying persons with unnatural life span or with strange geographical patterns (e.g., with only one transactions associated to a different parish that all the other transactions) can point out to interpretation errors in the transcription. However, as has already been demonstrated in the previous section, combining the data with the network structure is a much more powerful tool that can help surpass the most obvious

problems of transcription. (Keats-Rohan, 2007, p 171) states

> *Context is for us the all-important key to understanding the complex relationship between name and identity*

and the network offers a model that automatically relates context to individuals or transactions. This approach consists mainly in propagating information from one vertex to its neighbours.

It appears that individuals with high betweenness and low degree identified in the previous section have generally a quite large range of temporal activity. For instance, the chapter of Cahors is involved in transactions from 1277 through to 1472. While this is not surprising for a corporation, such a long lifespan is impossible for real persons. For instance, 'Arnaud Escairac' appears in only 10 transactions but these date from 1333 through to 1481: this is clearly an error of interpretation in during the transcription and the name probably corresponds to several namesakes. Indeed, as shown in Figure 5, Arnaud Escairac appears mostly in transactions around 1479 with persons that appear in other transactions with compatible dates. However, he also appears in two transactions from 1333 which involve persons who appears in other transactions with dates compatible with 1333. Then, the two sub-networks (the one around 1333 and the one around 1479) seem to be consistent and the only reasonable explanation is that the name 'Arnaud Escairac' corresponds to at least two distinct individuals living in two different centuries.

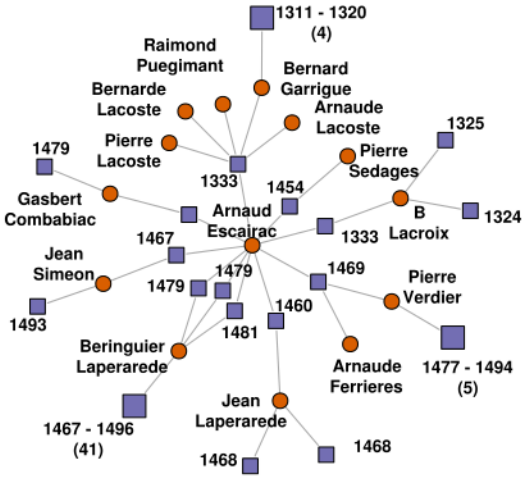

Figure 5: Local network neighbourhood for 'Arnaud Escairac' up to the third neighbour. Squares correspond to transactions, larger squares summarize a set of several transactions (the number of transactions is given between parentheses) and circles correspond to individuals. Transaction dates are given on the transaction nodes when they are known (the time period spanned by several transactions is given by an interval).

Obvious cases such as this one can be handled easily, the network analysis acting here only as a convenient means of detection. Notice however that Arnaud Escairac was already singled out during the analysis of the suspicious connection between the 'Ratier' cluster and a cluster whose central individual is named

Jean Laperarede (see last paragraph of Section 3): all direct shortest paths between the two leaders of the clusters go through Arnaud Escairac. It turns out that Arnaud Escairac is responsible for the direct connection between those clusters: if its vertex is removed from the graph, there is no longer a shortest path between Jean Laperarede and Ratier that goes directly from one cluster to the other one. This explains Arnaud Escairac's large betweenness and also the direct connections between those two clusters: due to the 'Arnaud Escairac' issue, a few persons have been attracted into the 'Jean Laperarede' cluster who should be in the 'Ratier' one, based on dates.

Arnaud Escairac was identified easily as a interpretation error during the transcription because of the unrealistically long lifespan implied by the confusion of the two individuals, but information propagation and information consistency principles allow more complex studies and give a clue to more subtle cases. Let us consider the case of Guiral Combe, one of the top betweenness individuals who does not appear in other rankings. He appears in 5 transactions from 1318 to 1370, a long but possible lifespan. A local network analysis is useful to get a better insight on this person. In Figure 6, the local network around this individual is extracted: it is the sub-graph whose vertices can be reached from Guiral Combe passing through 2 edges at most. Additional information is provided on this representation: the individuals' names, for vertices corresponding to individuals, the dates and the parishes for vertices corresponding to transactions. Using this information, two distinct groups of transactions clearly appear: the first one contains four transactions, all related to a place located in the parish named Capnié and all carried out with people from the Laperarede family between 1345 and 1370. The second group contains only one transaction, from 1318, with Guilhem Bernard de Prestis over a place located in the parish of Saint-Sernin. This information and the fact that Saint-Sernin does not border Capnié seem strongly to indicate that this vertex has also assimilated two namesakes. A deeper historical analysis, returning to the source material, would probably confirm that assumption but this example already shows how an automatic network analysis can stress interpretation issues in the transcription that would have been hard to find out without such tools. Direct transcription errors (such as, e.g., error in a date) could also be retrieved with a similar analysis.

6 Conclusion

This chapter has presented a network model and associated data mining tools for the exploration of a large database built from a corpus of medieval notarial acts. This model is general enough to be used in a wide variety of problems where entities are linked to each others by one or several types of relations. Although the model used in this paper is a simple relational model, additional attributes might be added to qualify the vertices (e.g. names, dates, places...) or the edges (e.g. transaction types) to describe the entities and the relations in a very precise way.

This chapter has rested on data mining tools dedicated to graph analysis, which include visualisation facilities, vertex clustering, numerical indicator calculation or local network extraction. Visualisation and clustering provided a relevant representation of our graph of notarial acts, which can be represented

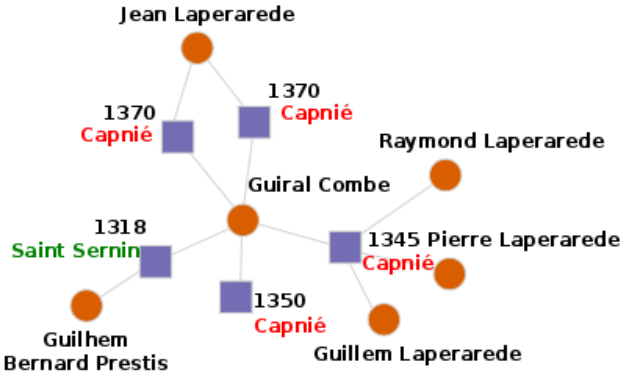

Figure 6: Local network neighbourhood for 'Guiral Combe' up to the second neighbour. Squares correspond to transactions and circles to individuals. The parishes' names are given in red and green and transaction dates are given in black.

either completely or in a simplified form to help the human eye understand the organization of its relations. Numerical indicators and zooming on a precise sub-network can be used to automatically select important individuals and also may make it possible automatically to identify and thus solve possible transcription errors or interpretation errors in the transcription that would not otherwise have been found. Using an interactive graph mining program, such as Gephi, these tools can be made accessible to researchers from other fields than computer science and mathematics. The next step would be to use this approach to correct the database, and then to run the analysis again to investigate any change. However, as this work is time consuming and demands the input of several well-trained persons, it remains a work in progress.

Finally, as astutely noted by one of the referees, an important finding of this work is that the network model is useful in deciphering transcription errors because certain of its characteristics (such as node centralities) are particularly sensitive to corpus bias or to errors in individual data. This fact must be kept in mind when using such models to avoid misleading conclusions.

References

Barabási, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

Barber, M. (2007). Modularity and community detection in bipartite networks. *Physical Review E*, 76:066102.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In Adar, E. e. a., editor, *Proceedings of the Third International AAAI Conference on Weblogs and Social Media*, pages 361–362. Menlo Park: AAAI Press, 2009.

Bertrand, M., Lemercier, C., and Guzzi-Heeb, S. (2011). Introduction : où en

est l'analyse de réseaux en histoire ? / introducción al análisis de redes e historia: herramientas, aproximaciones, problemas. *Redes*, 21(5).

Dorogovtsev, S. and Mendes, J. (2006). *Evolution of Networks*. Oxford University Press, Oxford, UK.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486:75–174.

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

Hautefeuille, F. e. a. (2009). Graph-comp: études des réseaux sociaux de sociabilité paysans au moyen-âge dans la châtelleni castelnau-montratier.

Kannan, R., Tetali, P., and Vempala, S. (1999). Simple Markov-chain algorithms for generating bipartite graphs and tournaments. *Random Structures and Algorithms*, 14:293–308.

Keats-Rohan, K. (2007). *Prosopography Approaches and Applications: A Handbook*, volume 13 of *Occasional publications of the Oxford Unit for Prosopographical Research*. Linacre College Unit for Prosopographical Research, Oxford, UK.

Lemercier, C. (2012). Formale methoden des netzwerkanalyse in den geschichtswissenschaften: warum und wie? österreichische zeitschrift für geschichtswissenschaften. *Austrian Journal of Historical Studies*, ÖZG 23:16–41.

Miquel, G. and Willy L., e. n. (2011). Archives départementales du lot. Technical report.

Rosé, I. (2011). Reconstrucción, representación gráfica y análisis de las redes de poder en la alta edad media. aproximación a las prácticas sociales de la aristocracia a partir del ejemplo de odón de cluny († 942). *Redes*, 21(1).

Rossi, F. and Villa-Vialaneix, N. (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.

Schaeffer, S. (2007). Graph clustering. *Computer Science Review*, 1(1):27–64.

Scott, J. (2000). *Social Network Analysis: A Handbook*. Sage Publications Ltd.

Tunkelang, D. (1999). *A Numerical Optimization Approach to General Graph Drawing*. PhD thesis, School of Computer Science, Carnegie Mellon University. CMU-CS-98-189.

Voloshin, V. (2009). *Introduction to Graph Theory*. Nova Science Publishers.