Reconstructing Medieval Social Networks from English and Latin Charters

Arno Knobbe
Marvin Meeng
Joaquin Vanschoren
Leiden Institute of Advanced Computer Science
Leiden University
the Netherlands

Sarah Rees Jones
Stefania Merlo Perring
Department of History
University of York
United Kingdom

1. Introduction

Over the past two years, the ChartEx\footnote{http://www.chartex.org} project has been developing new ways of analysing historical documents in an integrated fashion, and reconstructing medieval social networks based on this analysis. Specifically, the project’s aim was to develop tools to deal with medieval charters: records of legal transactions of property of all kinds: houses, workshops, fields and meadows. The charters also describe the people who lived there, and their relation to others. Long before records such as censuses or birth registers existed, charters were and still are the major resource for researching people, for tracing changes in communities over time and for finding ancestors. We are working with several collections of charters from the 12th to the 16th century, predominantly from England, both in the form of recent translations to English, as well as in the original Latin form.

The project was started by an international consortium under the Digging into Data\footnote{http://www.diggingintodata.org} program, bringing together historians and experts in Natural Language Processing, Data Mining and Human Computer Interaction. Partners from the following countries were involved: Canada, United Kingdom, United States, and the Netherlands. The role of the NLP experts was to automatically annotate the large collections of charters for further processing, based on example annotations of a sample of charters produced by the historians. The data miners then proceeded to link charters based on the actors and sites mentioned, and thus reconstruct some of the social network embedded in the collections. This will be the main focus of this paper. Finally, the HCI component of the project focused on the effective communication of found results through a Chartex Workbench.

2. Charters

The charters in our collection record transactions of property. Typically, they contain fairly concise descriptions of the grantor and recipient of the transaction, some description of the
property involved, often made more precise by mentioning the previous owners, and finally a list of witnesses. The following is an example of a charter from the Vicars Choral collection (pertaining to properties in the York, UK area):

408. Grant by Thomas son of Josce goldsmith and citizen of York to his younger son Jeremy of half his land lying in length from Petergate at the churchyard of St. Peter to houses of the prebend of Ampleford and in breadth from Steyngate to land which mag. Simon de Evesham inhabited; paying Thomas and his heirs 1d. or [a pair of] white gloves worth 1 d. at Christmas. Warranty. Seal.
Witnesses: Geoffrey Gunwar, William de Gerford[by], chaplains, Robert de Farnham, Robert le Spicer, John le plastrer, Walter de Alna goldsmith, Nicholas Page, Thomas talliator, Hugh le bedel, John de Glouc’, clerks, and others.
January 1252 [1252/3]

Note that this charter primarily identifies two people, *Thomas, son of Josce, goldsmith and citizen of York* and *Jeremy*, his younger son. The other person mentioned in the body text is *mag. Simon de Evesham* (mag. for magister, an academic title at the time) does not play a direct role, but is mentioned to specify a piece of land that is needed to identify the specific property being transferred here. The transaction relates to half of the land previously belonging to Thomas the goldsmith, and is mostly identified by how it is positioned in relation to other properties or landmarks, that were perhaps easier to identify at the time. Note that Petergate and Steynate are crossing streets that still exist in York as Petergate and Stonegate. The list of witnesses offers some clue as to the people involved, but does not play a major role in our work. Finally, the document is dated fairly accurately, but this is definitely not always the case, and differs from collection to collection, becoming common only after c. 1300.

One important thing to note here, that plays an important role in the process of record linkage, is the fact that most information conveyed in a charter is in natural language, something which hinders direct interpretation of the documents and leaves room for ambiguity. For example, one could argue in this text that Josce is in fact the goldsmith and citizen of York, rather than Thomas. Additionally, there is no notion of registered land or geographic coordinates, nor do people have social security numbers, as one would expect in modern legal transactions. To make matters worse, there was no unified spelling of people and place names, such that a considerable level of flexibility will have to be assumed when matching names across documents. Also, the notion of last names was only slowly appearing in Medieval England, such that people often are only identified by their first name. In many cases, people’s origin or profession served as last name, such as with William de Gerfordby or Robert le Spicer, but these ‘last names’ did not serve as family names. Needless to say, the unequivocal matching of people and sites across charters is a challenge.

3. Record Linkage

Despite the hurdles mentioned in the previous section, linking people and sites across charters turns out to be quite doable. For one, the population at the time was much smaller, and the people involved in property transactions is only a fraction of that, since not many people could afford to own land. Additionally, the charters tend to provide sufficient ‘cir-
cumstantial evidence’ in order to recognize people, perhaps as an unconscious attempt of the author to be sufficiently specific. For the charter mentioned above, *Tomas son of Josce goldsmith and citizen of York* is actually a phrase that appears in another charter, as well as the son Jeremy, although it appears both as *Jeremy* as well as *Jeremias* there. This charter also mentions that Jeremy is the son of Mariot, who is (by then) the widow of Thomas. This demonstrates how a social network slowly appears when being able to link individuals across charters. The links in this network can represent family relations explicitly mentioned in the document, but also the property transactions themselves: a charter connects the grantor and recipient. In the networks presented, we also include *might-be* connections, such that we can communicate persons mentioned in multiple charters, with various degrees of certainty.

Our record linkage method combines a probabilistic approach with a certain level of logical reasoning. The probabilistic side of our method aims to determine whether a candidate link (two mentions in two charters refer to the same person) is very probable, given the evidence available in both charters. Generally, for each person (a subset of) the following information is available: first name, occupation, title, last name, family relations. For each matching item (ignoring the complications of spelling variation for now) between the two persons, we need to determine the probability of making a wrong assumption of identity, and combine these probabilities in an overall confidence score of the assertion that we are dealing with one and the same person. Obviously, the more pieces of evidence we have and the more reliable that evidence is, the higher our estimated confidence.

One of the big challenges here is to estimate probabilities for individual items. For example, finding a Thomas in two charters may not be much evidence, if Thomas is a very common name at the time. Therefore, we need to estimate the frequencies of all first names in order to compute the partial probability. We opted to do this in ChartEx simply by using the combined collections as a source of names statistics, producing a histogram of all names appear in the collections. The five most common names found here, in descending order, are John, Thomas (unfortunately), Robert, William, Richard. Josce appears only twice, making the match in the previously-mentioned charters much more probable than the one concerning Thomas. The same process can be repeated for the occupations (yeoman, gentleman, esquire, clerk, goldsmith, ...). For family relations, one can simply adopt the same reasoning as for first names. Knowing one’s father is called Josce is just as informative as being called Josce oneself, so we can simply use the first name statistics. The process of determining the probabilistic contribution of a last name was less clear, for reasons mentioned earlier. For the lack of reliable statistics on last name occurrences, we simply introduced a fixed score, such that a matching last name contributes to the confidence by a constant amount.

Aside from the probabilistic reasoning described here, there was also a considerable amount of logical reasoning, notably when conflicting evidence was present. For example, having a mother with a different name is problematic, regardless of the other matching evidence. The same is true for appearing in two charters that are separated by more than a hundred years, although this reasoning is less clear-cut when the separation reduces to say 30 years. For lack of a good model of longevity in the Middle Ages, and more importantly in what age bracket one might be expected to be involved in property transactions, we introduced a fairly simplistic probability function depending on the number of years between
the charters, with all separations over 80 years being discarded. Some more sophisticated reasoning, that reflects some of the logic of charters and of the era, was involved when considering when considering relations that potentially do not last, or change over time. For example being married to person A doesn’t exclude one from being married to another person B at a later stage, or being one’s widow. As an aside, expressions such as son of and mother of were used to infer some of the genders of lesser-known names such as Fange (male) and Thomasina (female).

Although the probabilistic approach taken produces very satisfactory results, replacing considerable manual labour by historians, it has a few drawbacks that it shares with many probabilistic approaches to record linkage. First of all, producing name and occupation statistics from the collections themselves introduces a certain bias, for the simple reasons that people may appear more than once. Especially the names of those who own a lot of property (e.g. Simon de Evesham) will appear higher in the ranking than is realistic, with the undesirable side effect that their matching actually becomes less confident. Sometimes, assumptions need to be made that are not supported by sufficient data, such as in the case of last names. Finally, a common complaint of probabilistic approaches is that the combined estimate of confidence assumes that the individual probabilities are independent, which they are often not. For example, first and last name frequencies are known to be quite dependent, although this example doesn’t apply to our data. It does however apply to first names and occupations, which are not independently distributed. Despite some of these drawbacks, the method appears to work sufficiently well, and as long as one doesn’t interpret confidences as absolute numbers, but rather as rankings, the confidence numbers are very usable.

4. Quantitive details

In total, five collections of charters were available to the project, being:

- The Vicars Choral (University of York), 125 charters manually annotated, English, 5,000 charters (dated).
- Borthwick (Borthwick Institute, University of York), 55 charters manually annotated, English.
- DEEDS (University of Toronto), 49 charters manually annotated, Latin, over 10,000 charters.
- Wards2 (The National Archives, UK), 48 charters manually annotated, English, 7,000 charters.
- Cluny (University of Columbia), 50 charters manually annotated, Latin, over 5,000 charters (dated).

In the English documents, a total of 112 different first names occur, where we assume different spellings are different names. Of these names, the gender of over 85% could be inferred from the context in which they appeared (for example Thomas, son of Josce implies Thomas is a male name). Of the names for which the gender was resolved, 36% was female.
It should be noted though that in absolute sense, women were much less mentioned than men, especially where ownership of property is concerned. In a ranking of names according to their frequency, the first female name (Margaret) appears at rank 15. Also, the common name John is over 17 times more common than Margaret. This Medieval gender difference is also indicated by the occupation statistics, where the first clear female ‘occupation’ (an annotation that was used somewhat liberally in this project) is ‘widow’ at rank 12, after clearly male occupations such as ‘yeoman’ and ‘esquire’.

5. Outlook

The ChartEx project has by now finished. The funding for ChartEx was relatively short, making all steps in the process somewhat proof-of-concept. Still, a working system was produced that allows historians to work with large collections of charters in an integrated manner. The record linkage activities continue in a new, somewhat larger project with The National Archives (UK), where not only medieval records are involved but also more modern civil records. There are also well-developed ideas for a follow-up project to ChartEx, called The Medieval Mine, which aims to exploit the new capabilities of analysing collections in their entirety, and mining the structured result by means of modern Data Mining techniques.

Acknowledgments

This research was supported by the Digging into Data program, funded among others by JISC, AHRC and NWO.
Figure 2: Partial network generated from a subset of charters. The ovals roughly cover the seven charters involved. Gray lines indicate hypothesised links between people in various roles in the charters. Note some of the spelling variations.