Social Networks 69 (2022) 123–135

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Networks from archives: Reconstructing networks of official correspondence in the early modern Portuguese empire

Agata Błoch[a,*,1], Demival Vasques Filho[b,1], Michał Bojanowski[c,1]

[a] The Tadeusz Manteuffel Institute of History, Polish Academy of Sciences, Poland
[b] Leibniz Institute of European History, Germany
[c] Kozminski University, Poland

ARTICLE INFO

Keywords:
Historical network research
Natural language processing
Portuguese empire
Digital humanities
Colonial studies
Correspondence networks
Early modern letters

ABSTRACT

Historical archives provide invaluable insights into societies of the past, including social networks. However, the required amount of traditional archival work makes historical network studies usually small-scaled. We consider the problem of processing a large corpus of unstructured textual information to extract network data. The corpus consists of almost 170,000 documents of administrative correspondence of the Portuguese Empire, from 1610 to 1833, catalogued in the Portuguese Overseas Archives of Lisbon. Our contribution is twofold: the method and the result. Firstly, grounded in the review of manual, semi-manual and automatic methods of network data extraction from natural language corpora, we propose and demonstrate an approach using modern natural language processing algorithms. This approach tries to mimic traditional archivist's coding practices and is applicable to large corpora of texts, for which manual coding is infeasible because of scale. We believe our approach is generic and adaptable to other substantive contexts, languages, and types of historical archives. Secondly, the dataset created is rich in additional information such as occupation, administrative affiliation, and geographical location of senders and recipients. We provide a preliminary network analysis suggesting that the dataset is an attractive material for historians and social network researchers for addressing research questions about the political and social evolution of the early modern Portuguese Empire, spanning the reign of seven Portuguese monarchs.

1. Introduction

The history of the early modern Portuguese colonial empire has been debated over the last decades in the Luso-Brazilian academia. The characterization of the empire as based on the political dependence of the peripheries on the metropole, widely propagated in the historiography of the second half of the twentieth century (Prado, 1957/1967, Viotti da Costa, 1997/2000, Fernandes, 1969; Novais, 1995), is currently contested by the creators of the theory of multi-continental monarchy (Fragoso and Gouvêa, 2009; Guedes, 2013; Fragoso et al., 2017). This theory emphasizes the role of different political and administrative networks within the overseas empire and stresses the agreements made by the local nobility representing municipal chambers with the Crown. The dialogue, negotiation, and competition provided the basis for those relations (Fragoso and Jucá de Sampaio, 2012). For advancing the debate on the multi-continental monarchy, it is important to identify not only the key political, economic, and social actors, but also the relationships between them and how the networks of these relationships changed over the years.

We believe that, among other things, the structure of these networks can reflect significant macro-level processes that took place during the seventeenth, eighteenth, and nineteenth centuries in the Portuguese empire, such as the process of colonies becoming more and more independent politically and economically from the Crown. Reconstructing relations between key actors of the Iberian centre and its colonies will enrich the current image of the early modern Atlantic world and broaden our understanding of the global Portuguese empire as a networked society where its structure might explain certain behaviours and macro-level tendencies (Xavier et al., 2016, 2018). Hence, we hope that future research based on the data we describe in this article will provide a more in-depth insight into the Atlantic world history, better explore the transoceanic connections, and provide a better understanding of the

* Corresponding author.
E-mail addresses: agata.natalia.bloch@uw.edu.pl (A. Błoch), vasquesfilho@ieg-mainz.de (D. Vasques Filho), mbojanowski@kozminski.edu.pl (M. Bojanowski).

[1] All authors have contributed equally.

https://doi.org/10.1016/j.socnet.2020.08.008

Available online 24 September 2020

socio-political consequences of the multi-continental monarchy.

To contribute to this debate, we discuss here the reconstruction of networks of administrative correspondence based on an extensive dataset from the Historical Overseas Archives of Lisbon. Currently, it consists of a corpus of almost 170 thousand documents of the administrative correspondence exchanged in the period from 1610 to 1833 between Portugal and its Atlantic ex-colonies, such as Angola, Brazil, Cape Verde, Guinee, and Saint Tome and Principe. We define, however, the timeframe of our analysis in this article as 1642−1822. This is the period from the establishment of the Overseas Council in Lisbon (1642), which then assumed the administrative supervision of all the colonies, until the declaration of independence of Brazil in 1822.

Our overarching goal in this paper is twofold. First, we describe how we used natural language processing techniques to identify and extract structured data from an otherwise unstructured data of free text. We extracted information such as the sender and the recipient of these documents, actors' occupation in the administration (e.g. king, governor, secretary, etc.) and in society in general (earl, marquis, priest, and so on), their gender, their institutional affiliation, and their geographical location. Thus, we turned this archival data into relational (network) data. We believe that the methodology presented here, although applied to particular historical sources in Portuguese, can be valuable to other researchers with similar challenges and helpful in avoiding the strenuous work of manually coding extensive archival data.

Second, by presenting selected statistics and graphs, we discuss how the dataset we created can help in answering various types of research questions historians and social scientists might pose about the history and politics of the modern Portuguese empire. As we describe below, this has significant consequences for the ways we process the data. In particular, we have explicitly decided to process and analyze the data in ways to avoid loss of information or detail. We believe that any forms of data selection or filtering should be guided by a precise research question one might have. The preview of network descriptives that we show here is a first step towards answering some of our research questions. However, it is not closely tied to any particular question and shows nearly the entirety of the archival data with which we are dealing. Through network analysis using this dataset, one can, among other things, identify the main actors of the early modern Portuguese empire under the Brigantine Dynasty, their relation with top-ranked officials, as well as the participation of minorities in the decision-making processes in the empire.

The remainder of the article is structured as follows. Section 2 covers relevant studies in building social networks from historical texts, the techniques normally used to collect network data from unstructured sources and the methods to construct node sets. In Section 3, we provide more details of the corpus which constitutes the source material of our work. We explain our methodology in Section 4, with a step-by-step discussion on how we use natural language processing, including a machine learning approach, to identify and tag significant entities in the texts, such as persons, institutions, locations, and more. This allows representing each document as a structured object containing all the extracted information in a more standardized form. Section 5 presents the networks we reconstructed from our data and the fundamental characteristics of these correspondence networks. The paper is concluded with Section 6, in which we discuss our experiences of processing unstructured data into structured network data.

2. Building social networks from historical texts

As we stated in the previous section, the primary goal of this project is to turn a vast amount of unstructured textual data from archives into network data. The use of network analysis for historical research has proved to be fruitful and is becoming increasingly popular for different historical periods. On some fundamental level, building network data based on archival sources or corpora of historical documents shares the basic principles with the process of designing any network study in general. This process involves addressing the issues of network boundary specification (Laumann et al., 1989) related to identifying what social entities are of interest (the nodes), what types of relations are relevant and how to "measure" them (the edges), and what attributes of nodes and edges are of interest (see also Adams, 2019). In contrast to more "active" modes of data collection, such as experiments or surveys, working with archival information may seem, on the one hand, more "passive" in the sense that the researcher does not have the freedom of interrogating the social reality with the tools he or she would like. The available information is limited to what the archivists, in the distant past or recently, have decided to record or on which facts and events a chronicler or historian decided to focus. Consequently, some of the choices concerning boundary specification have been already made. Nevertheless, quite some freedom is often left to the network researcher in interpreting and coding the source material to infer, for instance, types of relationships and attributes of actors (see also Lee and Martin, 2015, including discussed references). The degree to which this is the case varies as much as the sources vary from structured indexes, catalogues or censuses to unstructured natural language prose of letters, chronicles or history books. While the former are often almost ready-made for quantitative analysis, the latter need to be carefully read and the network information extracted. It is worthwhile to sketch that continuum with examples of existing historical network studies to position the presented work among them.

Studies of personal lending in fifteenth-century Florence (Gondal and McLean, 2013b, 2013a; McLean and Gondal, 2014) are good examples of a network analysis of a structured historical source. The data came from originally hand-written maps and tables of Florentine *Catasto* from 1427, which have been already turned into a database by historians (Herlihy et al., 2002; Herlihy and Klapisch-Zuber, 1985). Still, network boundary choices also had to be made to restrict attention to the explicitly defined elite families and significantly high financial obligations (McLean and Gondal, 2014, p. 144). Other historical network studies that benefited from digitization of otherwise structured historical information include research on bipartite networks arising from medieval texts contained in multiple medieval manuscripts (Riva, 2019; Vozár, 2018) – the input for the network analysis were digitized tables of contents of a large number of medieval volumes. Our source material is a corpus of an unstructured text with very little metadata; thus, we need to look into studies that build networks out of natural language prose.

Techniques of assembling network data from unstructured text vary strongly depending on precise research questions posed. Studies pursuing questions about communication, discourse, discussions and narratives look into methods of representing text as an abstract network of interrelated phrases, concepts, descriptions of identities or emotions. Such content analysis (Franzosi, 2004, 2008; Popping, 2000, 2003; Popping and Roberts, 1997) can also be found in studies of historical sources. Examples include studies of proceedings of political bodies (Fuhse et al., 2019; Padgett et al., 2019), roles and practice of charity organizations (Mohr, 1994; Mohr and Duquenne, 1997), political discourse based on newspaper articles (Franzosi, 1997), parliamentary speeches (Curran et al., 2018), or other types of text messages (Alexander and Danowski, 1990; Danowski, 2009). From the perspective of our goals, we are primarily interested in identifying network nodes corresponding to physical social entities such as persons or organizations rather than abstract semantic concepts or meanings. As it will become more evident in Section 3, the texts in our corpus were written by contemporary historians and archivists and as such are not of particular interest from the perspective of narratives or mental models. Still, we do plan to employ some of these methods to capture the topics of our documents – an issue to which we go back in the discussion in Section 6.

2.1. *The art and sweat of working with prose*

Network studies using unstructured textual sources to reconstruct historical social networks between a relatively well-defined set of actors flourished starting with the pioneering research of Padgett and Ansell (1993). Based on a careful study of Kent (1978), as well as some other sources, they built a network dataset of nine different types of relations based on kinship, economic, political and friendship relations between the Florentine families. The study may seem small-scaled to a person who never personally did any archival research, which among its challenges has the (un)availability of documents and their scattering throughout several sources, and requires persistence of extracting exact information, especially important when studying small systems. Manual extraction of network data is an arduous task which requires reading all the sources (either archival sources or well-defined dataset and databases) and storing useful information in a digital format. Moreover, this process is not mechanistic but full of essential decisions consequential for the resulting network data. For example, Padgett and Ansell had to make network-boundary choices for which families to include out of the many familial lineages present in fifteenth-century Florence. The study of Baker and Faulkner (1993) is an example of applying similar manual techniques to a more contemporary corpus of testimonies before the U.S Senate. Other examples include studies of historical sagas or chronicles to reconstruct networks among the real or fictional characters from, among others, writings of Pericles (Cline, 2020), Petronius (Köstner, 2019), Tacit (Köstner, 2020), or Scandinavian sagas such as "Beowolf" (Kenna and Mac Carron, 2016; Mac Carron and Kenna, 2012, 2013). The authors of these studies had to manually build dictionaries of actors and decide upon the existence of edges linking them based on the texts and own judgement.

A special place among such intrinsically manual network studies takes research on correspondence, in particular those originating from the early modern period. A letter is a relatively standardized form of communication with a rather clear set of metadata: the sender, the recipient (usually including their geographical locations) and the date. Our corpus also includes letters next to other types of communication documents such as petitions. The manual processing of letter collections allowed studies of correspondence between protestant intellectuals (Ahnert and Ahnert, 2015), family and friends of Cicero (Gilles, 2020), Pliny the Younger (Germerodt, 2020), or politicians of the Late Roman Republic (Rosillo-López, 2020). Other correspondence network studies benefited from large scale digitization of culture projects such as the Republic of Letters (Edelstein et al., 2017; Edelstein and Kassabova, 2018), Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic (Roorda et al., 2010; Van Den Heuvel, 2015; Van Den Heuvel et al., 2016) or The Reception and Circulation of Early Modern Women's Writing, 1550–1700 (Booth et al., 2017; Coolahan, 2017).

It is important to realize the scale of some of the example research mentioned above. While Ahnert and Ahnert (2015) uses 289 letters, Bourke (2017) uses 4,708, McShane (2018) uses 405, and Gilles (2020) use 914, our sources comprehend almost 170,000 correspondence documents. It is not feasible to manually extract network data from them because manual work does not scale well with corpus size. Thus, we have to resort to computational methods to achieve the goals set for our project. As we will demonstrate in Section 4, our methodology also has a manual component. However, our manual component is of a different type from these above. Instead of extracting the network data itself, we annotate documents for feeding the machine learning algorithms that will, in turn, help the extraction of network data from the whole corpus.

2.2. Towards an AI archivist

Early applications of computational methods to network data construction from text corpora still resorted to manual work to arrive at the proper set of nodes, but took advantage of the corpus structure and available metadata to produce the edges of the network of interest. Examples include manually indexing persons mentioned in entries of biographical dictionaries to build a two-mode network of person-dictionary entry relations, with a subsequent projection step arriving at a unimodal network of persons (Warren et al., 2016a). In a mostly similar fashion, researchers had studied intellectual co-authorship of historical works (e.g. Breure and Heiberger, 2019). This technique has been framed as "distant corpus reading" in the digital humanities (Grandjean, 2016, 2017; Moretti, 2013). More elaborate techniques of defining co-occurrence have been developed, such as with a moving window of words (Carley, 1997; Carley and Palmquist, 1992). While they are not directly useful for our goal of identifying social entities, we are likely to reach for them in the future when focusing on relations between actors who are not senders or recipients of the documents in our corpus, but mentioned in the content of these letters.

The advent of named entity recognition (NER) algorithms permitted addressing the problem of scale when constructing node sets. A NER model tags parts of text corresponding to a named entity such as a person or a name of a location. Early applications of NER models (Magnini et al., 2002) were used as a preliminary step of constructing dictionaries of person names or thesauruses of more abstract concepts (Diesner et al., 2012; Diesner and Carley, 2005). In historical network research, they have been successfully used to, among others, reconstruct social networks based on the University of Luxembourg archives (CVCE), documenting the process of European integration in the period 1945–2009 (McGee et al., 2016), what now constitutes part of the HistoGraph project (Wieneke et al., 2014, 2013). Similarly, the LONSEA project (Sibille, 2011) assembled an interactively-explorable network dataset of persons, organizations and events based on documents collected in the United Nations Archives in Geneva. In our project, we very much follow their footsteps but use more modern natural language processing technologies and NER models that rely on artificial neural networks. However, among other challenges that we will discuss shortly, it turned out that pre-trained models for Portuguese provided poor performance in detecting historical first and family names as well as historical names of geographical locations. That led us to train a NER model ourselves by doing a traditional archivist manual work first (as we described above in Section 2.1), to provide its results as a set of examples for a machine to learn.

3. Archival research and description of archival sources

In this section, we discuss our source documents. The Historical Overseas Archives in Lisbon[2] stores a separate collection of the Overseas Council – the only institution of the early modern period with political, administrative and financial responsibilities with respect to the whole Portuguese Empire – that gathered documents produced by itself and its subordinate colonial institutions. These documents are individual letters preserved in paper form as thick binders, dating from 1610 to 1833, and represent the main existing administrative correspondence between the Portuguese metropole and its Atlantic overseas colonies, such as Brazil, Cape Verde, Guinea (-Bissau), and São Tomé and Príncipe Islands. The letters contain political and bureaucratic issues, and can reveal the day-by-day life of the colonial inhabitants (Bellotto, 2004).

In recent years, archivistas in several research projects made a massive work of cataloguing these correspondences: "Projeto Resgate"[3] catalogued the series of documents from Brazil; "Projeto África Atlântica"[4], from Angola, Guinea and São Tomé and Príncipe; and "Projeto

[2] The Historical Overseas Archives in Lisbon: https://actd.iict.pt/collection/actd:CU

[3] Projeto Resgate http://bndigital.bn.gov.br/dossies/projeto-resgate-barao-do-rio-branco/

[4] "África Atlântica: da documentação ao conhecimento, sécs. XVII-XIX" coordinated by Ana Isabel Silva Canas da Cunha Delgado Martins under research project no. TDC/HIS-HEC/103483/2008. Instituto de Investigação Científica Tropical,I.P. (IICT) (IICT/MNE). Available: https://www.fct.pt/apoios/projectos/consulta/aavaliar.phtml.pt?iep=4385

Resgate do Acervo Histórico de Cabo Verde em Portugal"[5], from Cape Verde.

By cataloguing, we mean that the archivists organised and separated the documents in collections following two criteria (Boletim, 1950): chronological, according to the date of the document; and geographical, according to the regional archive in which the document was originally stored – the geographical criterion does not indicate the sender's location when writing the letter.

The archivists created several registers for these collections (25 related to Brazil and 5 to the Western African region), explaining the two criteria used and containing thousands of entries apiece. Within a register, each of the thousands entries corresponds to an original letter and has a summary of the letter's content. The registers serve as a guide to researchers when doing archival research and are digitally available as unstructured (and unstandardized due to the several projects) PDF files. We used these files to extract network data, as we will explain shortly.

Despite having access to the digital registers, we had to conduct a thorough archival research. The heterogeneous nature of the records demanded research not only in the Overseas Archives, in the National Library and at university libraries in Lisbon but also in libraries in Brazil, for a better understanding of the administrative and political structures of the Portuguese Empire. The period we contemplate is very dynamic in terms of bureaucratic development of the Portuguese colonies. We had to be attentive to the changing policies, offices and even political centers, including the relocation of the Brazilian capital from Salvador da Bahia to Rio de Janeiro in 1763 and the transfer of the capital of the Portuguese Empire from Lisbon to Rio de Janeiro in 1808. Understanding these matters and the characteristics of the registers were crucial to the development of our computational methods.

Figs. 1 and 2 present an example of the original paper document (as stored in the Portuguese Overseas Archives in Lisbon) and the corresponding register entry, respectively. This letter comes from the *Avulsos* catalogue from São Tomé and Príncipe Islands. The summary in the register describes why the document was sent and contains information about the sender and recipient, their occupation, and date and place of dispatch. It lacks emotions or personal feelings when compared to the paper original. For example, the sender of the original (Fig. 1) is a noblewoman who states, among other things, that she "is dismayed, being a noble widow who lives recoiled, having nothing else that can be worth more than the fruits of her own farm, and reduced to the greatest misery for not being able to feed herself and her family". The corresponding summary (Fig. 2) only mentions that the sender is a "noblewoman who lived with difficulties".

The administrative correspondence we analyze reflects complex social and political processes of the empire and (as mentioned before) is an integral part of the official collections which the Overseas Council duly registered and stored. As an official corpus, it is not subject to the bias of private collections. Ahnert and Ahnert (2015) expressed such concern regarding the focus of Foxe's collection – their source material – on the correspondence of martyrs of the Protestant church, neglecting those written by other protestants leaders spared by the Catholic church.

Furthermore, the collection also avoids the bias generated by those trying to "employ archives to institutionalize their power" (Jimerson, 2003). Letters sent by private persons, such as slaves or Native Americans, were still stored in Lisbon, and any further exchanged correspondence was also copied and attached. Some original letters may be in family archives in Brazil or Africa, but a copy should be part of the available documents.

Yet, our source is still subject to the vicissitudes of time. It is impossible to determine how many documents could have been destroyed during the massive fire that devastated Lisbon in 1755 as a consequence of an earthquake. It is also difficult to estimate how many documents did not make their way to the archive and remained in the colonies. Unfortunately, limitations regarding the availability of historical documents are difficult (or even impossible) to overcome. As researchers dealing with this type of documents, we have to bear in mind such limitations when performing our analyses and drawing our conclusions. In our future work, however, we aim to investigate other sources and archives to complement any possible missing correspondence.

Similarly to the research papers examining the correspondence network discussed in Section 2, we consider the following elements essential to our dataset: date, type of document, place of origin, sender, recipient, content, and any third parties involved in the text. These elements are unavailable as metadata and thus need to be extracted from the content. At this stage, we are not concerned about carrying out a linguistics-based approach (Franzosi, 1998) nor about evaluating what type of relationship the sender maintained with the recipient (Kenna and Mac Carron, 2016). Our focus now is to identify the key social actors, their attributes, and the institutions they represent. It is in our research agenda, however, to address the semantic content of the documents to group the correspondence according to the subject matter, as the colonial representatives used to discuss each problem in separate letters.

Each register entry contains a specific date in the form of a day, month and year. Sometimes these dates were given as approximate or determined by the archivists, containing only the year. Fig. 3 shows the evolution of the number of documents per year, in the period of our analysis (1642–1822). The rise of transatlantic correspondence during the reign of John V reflects the changing policies of the Lisbon metropole, which previously focused its attention on the Portuguese state in India. However, the 1755 earthquake in Lisbon might have had a strong impact on decreasing the circulation of correspondence between the destroyed capital of the empire and its colonies, during the reign of the monarchs Joseph I and Maria I. After stabilizing, the number of records peaked under Maria I for about 10 years – a possible explanation is that this 10-year period precedes the fleeing of the Portuguese royal family and nobility to Brazil, due to the invasion of Portugal by Napoleon's forces. Investigating the relationship between the increased communication activity and the transfer of the Portuguese court to Brazil is another item for future work.

The archivists also classified the documents into over five hundred categories according to their type. The most frequent types are: Requerimento (Petition), Ofício (Circular), Carta (Letter), Consulta (Enquiry), and Aviso (Notice), as shown in Fig. 4. The classification is not always standardized, again due to the several parties involved. For example, we noticed that some seventeenth century documents labelled as Carta (Letter) reveal a similar writing pattern and content as in the eighteenth century documents labelled as Requerimento (Petition). Moreover, some documents are not really a correspondence per se, as they do not have a relational character with a defined sender and recipient. These documents usually refer to an administrative decision, such as Bilhete (Ticket), Passaporte (Passport) or contain only maps or tables.

The meticulous work of the members of the Overseas Council facilitated the task of determining the place where the documents were stored (Boletim, 1950). In the case of the Brazilian colony, the country was not perceived as one integrated territory, but initially as hereditary captaincies and then as states. The great challenge, however, was to identify the location of the sender, considering that the name of some places are no longer in use; other places do not exist anymore, like sugar cane plantations or villages. Generally speaking, the letters were usually sent from very different cities in the case of Brazil, and from major strategic cities, ports or fortalezas (fortresses) in the case of Africa. Fig. 5 presents the number of entries from every place of origin identified.

The senders of the letters are both secular (e.g. municipal chambers) or religious institutions (e.g. brotherhoods), as well as colonial officials (as governors and viceroy), clergy (bishops, priests), military personnel

[5] More information about "Projeto Resgate do Acervo Histórico de Cabo Verde em Portugal" are available on the website of the Overseas Archives in Lisbon: https://actd.iict.pt/collection/actd:CU.

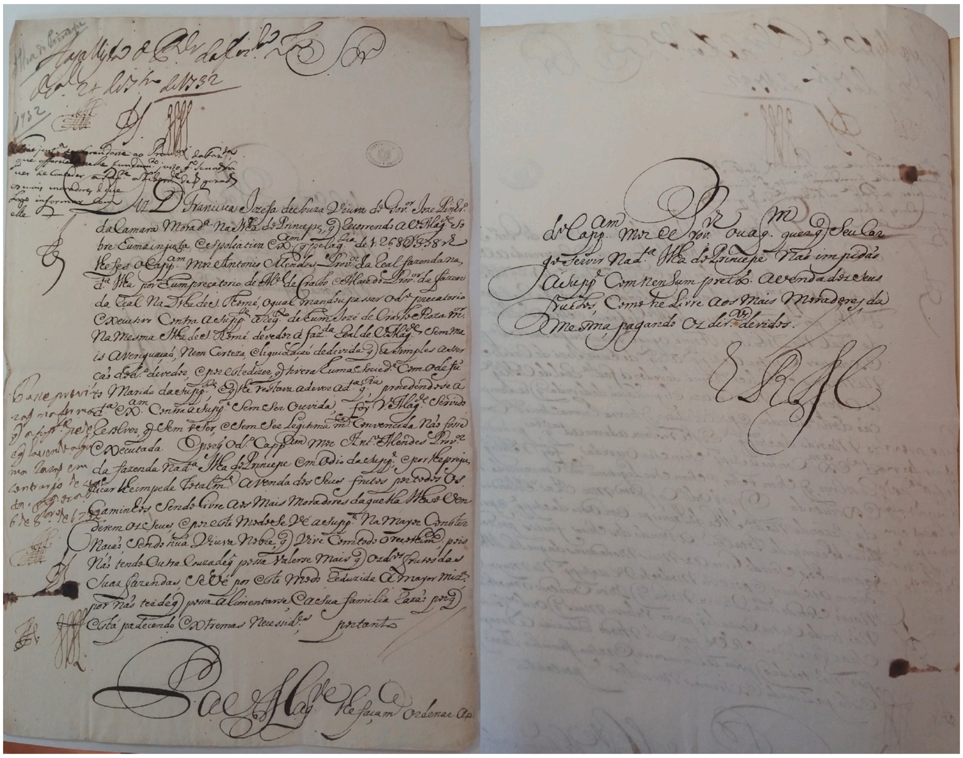
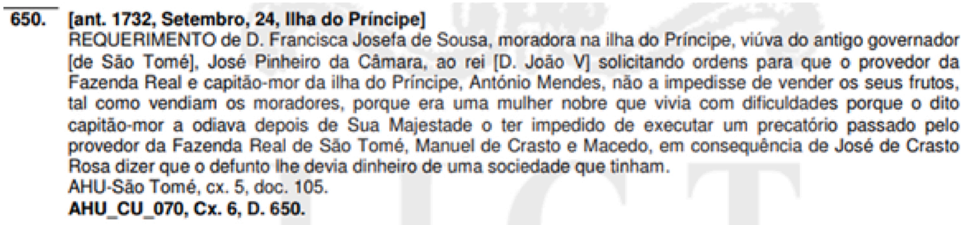


Fig. 1. Petition of Francisca Josefa de Sousa to King John V (1732). Source: Original documents from São Tomé and Príncipe Islands. Arquivo Histórico Ultramarino, CU_070, cx. 6, doc. 650.



650. **[ant. 1732, Setembro, 24, Ilha do Príncipe]**
REQUERIMENTO de D. Francisca Josefa de Sousa, moradora na ilha do Príncipe, viúva do antigo governador [de São Tomé], José Pinheiro da Câmara, ao rei [D. João V] solicitando ordens para que o provedor da Fazenda Real e capitão-mor da ilha do Príncipe, António Mendes, não a impedisse de vender os seus frutos, tal como vendiam os moradores, porque era uma mulher nobre que vivia com dificuldades porque o dito capitão-mor a odiava depois de Sua Majestade o ter impedido de executar um precatório passado pelo provedor da Fazenda Real de São Tomé, Manuel de Crasto e Macedo, em consequência de José de Crasto Rosa dizer que o defunto lhe devia dinheiro de uma sociedade que tinham.
AHU-São Tomé, cx. 5, doc. 105.
AHU_CU_070, Cx. 6, D. 650.

Fig. 2. The register entry corresponding to the petition of Francisca Josefa de Sousa to King John V shown in Fig. 1. The majority of the register entries follow the same pattern as the one pictured here: the sender and its attributes, the recipient and its attributes, and the summary of the content of the letter.

(soldiers, captains), or the individuals, both women and men, regardless of their social status (indigenous, slaves), or colour (black, “mulatto”). The most frequent recipients were those residing in Lisbon: the Portuguese monarchs, the Overseas Council and the secretaries of state. However, many letters represented local communication, sent to the governors of a colony or a state, and this type of communication was very characteristic for Mozambique.

There is a substantial diversity in topics that were the subject of letters sent to and from Portugal. The ones sent from the overseas possessions often contain information on the social, administrative, political, religious, and economic character. The inhabitants of the Atlantic colonies – including indigenous, fortune hunters *Bandeirantes*, intermediaries *Pombeiros,* pirates – informed their recipients about the exotic products, culture, people, plants and animals; their memories as settlers; and their socioeconomic and political situation. The colonial settlers often petitioned, hoping for a favorable solution to their problems, asking the king to provide them with self-defense weapons or complaining about the inefficiency of the system. In turn, the Lisbon administration used to inform them in the form of instructions named *Carta Régia* (Royal Charter), *Regimento* (Regulation), *Lei* (Law), *Provisões* (Provisions), *Consultas* (Consultation), and *Instruções* (Instructions) that addressed the politics, economy or the secular and religious power structures. The colonial residents were also instructed on how to wage war against the natives, or how to plant sugar cane, build fortresses, maintain the peace, buy slaves, run transatlantic trade or extract precious stones from the mines (Boschi, 2011).

4. Turning archival data into network data

In this section, we discuss the course of action for developing the digital methodology we used in the process of turning the archival data into network data. Fig. 6 shows a diagram with a summary of the steps of our methodology. In a nutshell, these steps are:

1 Random sampling: Creation of a random sample from the 169,221 register entries in plain text, ready for annotation.
2 Text annotation: Annotation of entries in the sample, with the labels (categories) of our interest.
3 NER extracts senders, recipients, and attributes: NER (Named Entity Recognition) model training with the annotated entries and parsing to identify the entities present in all entries.
4 Regex: Identification of text patterns with regular expressions.
5 Metadata extraction: Identification and extraction of metadata included in the text, also using regular expressions.

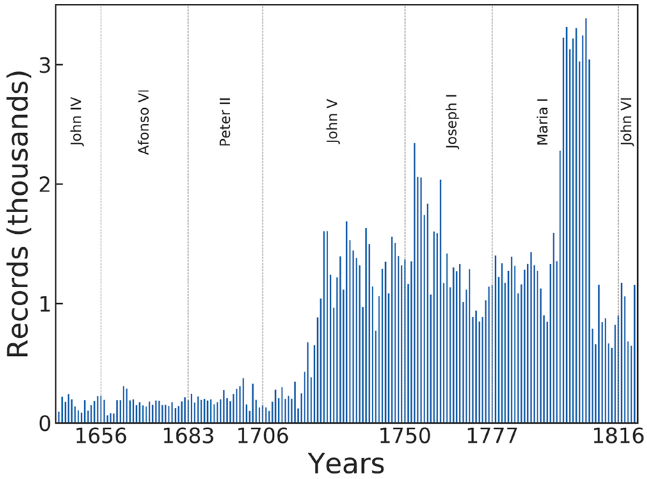


Fig. 3. Evolution of the number of records in the period chosen for our analysis (1642-1822). Gray dotted lines correspond to years in the x-axis, when there was a transition of monarchs. The number of records during John V might have increased due to the shift of attention of the metropole from India to the Atlantic colonies. The peak during the second half of Maria's reign might be related to the Portuguese invasion by Napoleon Bonaparte.

6 Network construction: Creation of network data after analysis of duplicate entities and correction of typos, and network analysis and visualization.

We will discuss these steps in more detail in the next two subsections and Section 5.

4.1. Identifying the entities

We have seen in the previous section that the correspondence entries of the Portuguese Overseas Archive collection have several distinct forms as well as a heterogeneous set of actors (senders and recipients). We identified these actors, and most of their attributes, using the named entity recognition (NER) tool of the natural language processing (NLP) package for *Python* called *spaCy*. The function of the NER tool is to parse plain texts to locate and identify the entities mentioned in such texts. Then, the tool classifies these entities into previously defined categories, e.g. people, institutions, locations, products, events, and so on.

Our first big challenge when using natural language processing techniques in this study was the fact that our texts are in Portuguese. Even though there exist libraries to perform NER in the Portuguese language, these libraries still are relatively small, without a large diversity of texts. As a consequence, although the reported accuracy for Portuguese reaches 89 %, for the spaCy v2.2 NLP library[6], in practice the accuracy can be much lower, especially for documents from the seventeenth until the beginning of the nineteenth century. Terms of such periods makes it more difficult for the NER analysis, as many names and locations are not frequently used anymore. Many of these names do not appear in the current text libraries used as a reference for the NER model.

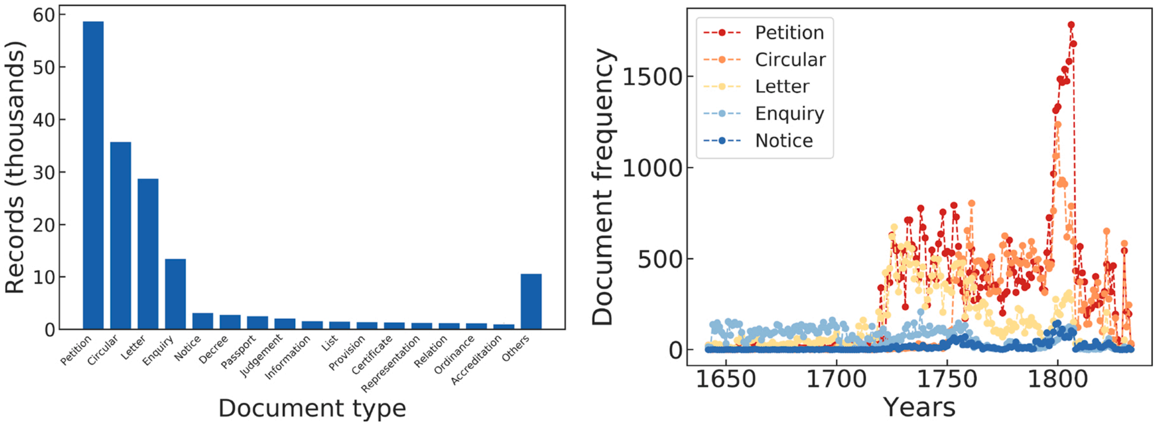


Fig. 4. Left: number of records per type. Right: evolution of the number of documents per type of the five most frequent document types. Petitions and circulars are the main type of document that resulted in the sudden rise of exchanged correspondence observed in the reign of Maria I.

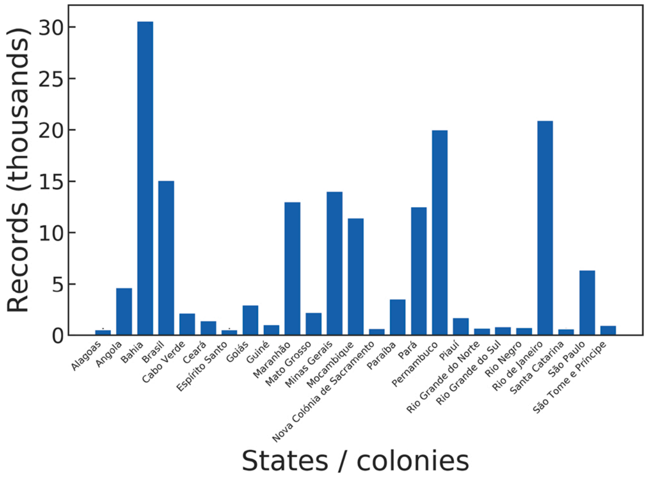


Fig. 5. Number of records for every place of origin (localization of the sender) identified.

Our second big challenge was that the variety of the correspondence types, their complexity, and the time-span made it difficult for the right categorization of the entities. The standard Portuguese NER model in spaCy has only four entity categories: *PER* (for persons), *LOC* (for locations), *ORG* (for institutions in general), and *MISC* (miscellaneous - for any other entity), which is not enough for our purposes. We wanted to enrich our network data with information such as gender, occupation, and nobility title, if any, of actors.

To overcome the issues mentioned above, we decided to train a new Portuguese model from scratch. For that, our first task was the random sampling. We created a random sample of 4230 entries – corresponding

[6] Information on SpaCy is available on https://spacy.io/ and information on SpaCy models is available on https://spacy.io/models/pt

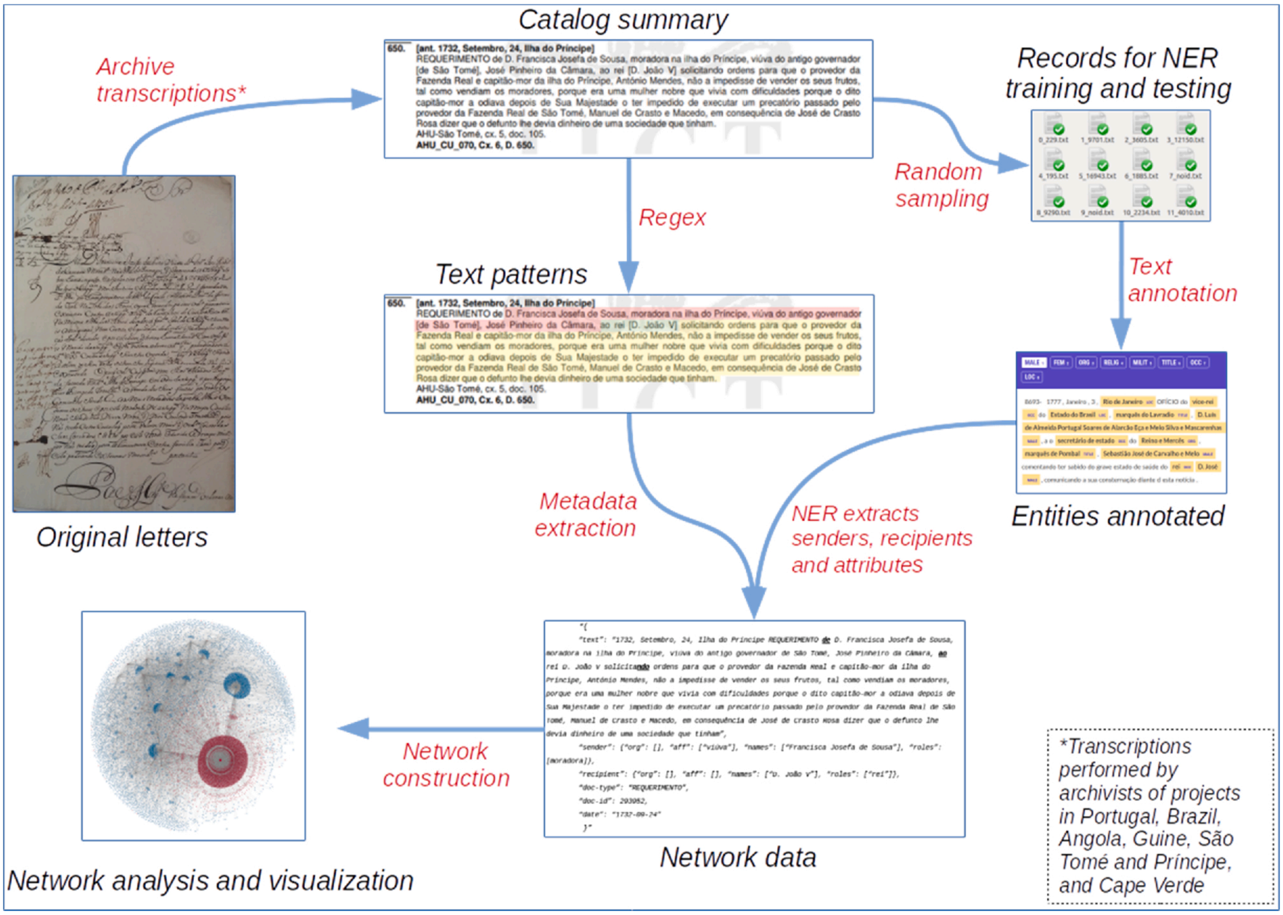
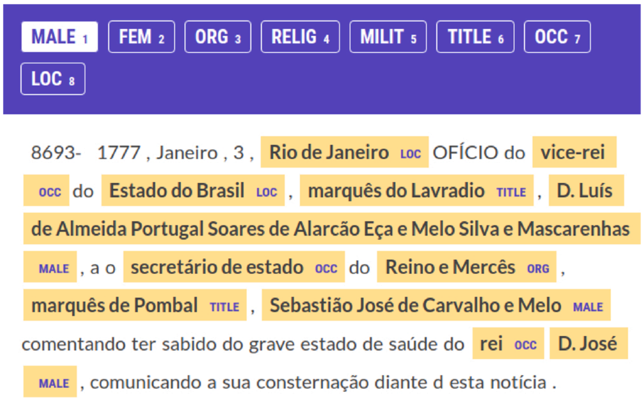

Catalog summary
Archive transcriptions*
Records for NER training and testing
Random sampling
Regex
Text patterns
Text annotation
Original letters
Metadata extraction
NER extracts senders, recipients and attributes
Entities annotated
Network construction
Network data
Network analysis and visualization
*Transcriptions performed by archivists of projects in Portugal, Brazil, Angola, Guine, São Tomé and Príncipe, and Cape Verde

Fig. 6. Schematic of the methodology to turn archival data into network data.

to 2.5 % of the total – to perform the second step: Text annotation. We manually annotated the entities and their type, for every entity found in these entries. To that end, we used the software Prodigy[7] . Fig. 7 shows an example register entry with manually marked text fragments referring to entities of our interest labeled with the proper categories.

We split *PER* into *MALE* and *FEM*. For *MALE* we decided to annotate all the complete male names, for instance *Antônio Teixeira de Mendonça*. For practical reasons, we decided that the abbreviation of the noble Portuguese title *Dom* should be annotated together with the name, as it was in the case of *D. João d'Ávila*. We performed analogically with *FEM*, by annotating such forms as *Ana Maria Nogueira*, *D. Joana or Dona Maria Braga*. To track nobility titles, we created a separate category named *TITLE*, including for instance *Marquis of Lavradio* and *Count of Oeiras*. For the categories of the administrative and political structure, due to such a diverse character of colonial entities in terms of linguistics, we decided to divide them into three categories. The first is *ORG*, which contains all civil and secular establishments of a purely administrative-political character, such as *Conselho Ultramarino* (Overseas Council) or *Casa de Suplicação* (Tribunal). We adopted the same practice in the categories *RELIG* and *MILIT*, for religious and military institutions, respectively. The category *OCC* includes all words relating to professional occupation, for instance *governor* (governor), *soldado* (soldier) or *juiz* (judge). Finally, for the category *LOC* we decided to annotate not only cities, states and countries, but also several locations like villages such as *Vila de Massangano*, rivers as *Rio São Francisco,* and military fortresses in Africa known as *presídios*, as *Presídio de Pedras Negras*. This way, we are able to identify more detailed information about the localization of the events described in the letters. We did not use the *MISC* category.



8693- 1777 , Janeiro , 3 , Rio de Janeiro LOC OFÍCIO do vice-rei OCC do Estado do Brasil LOC , marquês do Lavradio TITLE , D. Luís de Almeida Portugal Soares de Alarcão Eça e Melo Silva e Mascarenhas MALE , a o secretário de estado OCC do Reino e Mercês ORG , marquês de Pombal TITLE , Sebastião José de Carvalho e Melo MALE comentando ter sabido do grave estado de saúde do rei OCC D. José MALE , comunicando a sua consternação diante d esta notícia .

Fig. 7. Example of annotating a register entry with Prodigy. In this case we have LOC – Rio de Janeiro, and Brasil; OCC – vice-king, secretary of state, and king; TITLE – marquis of Lavradio, and marquis of Pombal; MALE – D. Luís de Almeida Portugal Soares de Alarcão Eça e Melo Silva Mascarenhas, Sebastião José de Carvalho e Melo, and D. José.

With over four thousand entries annotated, we enter the third step of our methods: NER extracts senders, recipients, and attributes. We used 80 % of all annotated entries to actually train the model and the other 20 % to test it. As a result of the test, we reached 93,1% of accuracy in recognizing the eight categories of entities.

[7] Information on Prodigy is available on https://prodi.gy/

4.2. *Extracting relational data and attributes*

After the identification and classification of the entities, our next step was to identify who of those were the sender(s) and the recipient(s) of the letters. In step number four, we had to distinguish possible patterns in the structure of the text of each correspondence entry that would characterise the activity of the entity as a sender or a recipient. In general, the register entries (the petitions in particular) present the following content segments: senders and their attributes (if existent), recipients also with their attributes (again, if existent), and a summary of the main content of the correspondence, as shown in Fig. 2.

That way, by using regular expressions techniques, we could divide the text of each letter into three segments, with the following reasoning. First, we located the presence of Portuguese prepositions such as *de, do, da, dos, das* (in English: *from* and *from the*) and set these words as the start of the text segment with information regarding the sender of the letter. Second, we located other prepositions as *a, à, ao, às, aos, para* (in English: *to* and *to the*), as the end of the sender segment and the start of the recipient segment. Finally, we located verbs in the gerund form (in Portuguese, verbs ending with *ndo*, in English: *ing*), indicating some activity as, for instance, *asking, claiming, protesting, stating, declaring*, and so on. The verbs in the gerund form are set as the end of the text segment related to the recipient, and the start of the segment with the content summary.

The fifth step, metadata extraction, consisted in extracting the metadata present in the text of the register entry, as shown in Fig. 2. Below is an example of a processed entry that we created, stored in JSON format, after the named entity recognition step, the identification of senders and recipients as well as their attributes, and extraction of metadata. This is the petition - the original and the record are shown in Figs. 1 and 2, respectively - sent by Francisca Josefa de Sousa, addressed to King John V. The characters marked in bold and underlined are the words used to identify the three segments of text as stated above. In *"aff"* we store the actors' affiliation to institutions, be it *ORG*, *MILIT* or *RELIG*.[8]

```
        "{
        "text": "1732, Setembro, 24, Ilha do Príncipe REQUERIMENTO de D. Francisca Josefa de
Sousa, moradora na ilha do Príncipe, viúva do antigo governador de São Tomé, José Pinheiro da
Câmara, ao rei D. João V solicitando ordens para que o provedor da Fazenda Real e capitão-mor
da ilha do Príncipe, António Mendes, não a impedisse de vender os seus frutos, tal como
vendiam os moradores, porque era uma mulher nobre que vivia com dificuldades porque o dito
capitão-mor a odiava depois de Sua Majestade o ter impedido de executar um precatório passado
pelo provedor da Fazenda Real de São Tomé, Manuel de Crasto e Macedo, em consequência de José
de Crasto Rosa dizer que o defunto lhe devia dinheiro de uma sociedade que tinham",
        "sender": {"aff": [], "title": [], "names": ["D. Francisca Josefa de Sousa"], "occ":
[]},
        "recipient": {"aff": [], "title": [], "names": ["D. João V"], "occ": ["rei"]},
        "doc-type": "REQUERIMENTO",
        "doc-id": 293952,
        "date": "1732-09-24"
        }"
```

With the above methodology we extracted senders and recipients. The senders are more numerous and diverse than the recipients. Among senders there are government officials but also less favoured individuals such as black slaves and indigenous representatives. While the recipients account for over 9000 actors, the number of senders amounts to over 44,000. The number of actors who are both senders and recipients is about 1,600.

At the beginning of step number six, network construction, we performed data quality assurance. Due to errors, like typos, found in the registers, we performed a thorough examination of the entities to avoid duplicates, especially of those more relevant to our analysis. That is, we ordered the entities by the frequency with which they appear in the records. Starting from the most frequent ones, we applied an algorithm that compares sequences of characters of a pair of elements (Ratcliff and Metzener, 1988). We used the most frequent version of each entity's name as the reference element and compared every other name in the sets of entities with that. We could fix, for example, instances like "Fernando José Portugal" to the corrected version "Fernando José de Portugal", and "Gomes Freire de Andrada" or "Gomos Freire de Andrada" to "Gomes Freire de Andrade". Thousands of duplicates were identified and fixed. It is worth noting that we did not apply the algorithm to automatically correct the names of monarchs. Otherwise, names such as "D. João IV" could easily be mistaken as "D. João VI", for instance. That would create the opposite issue of duplicates. Instead of treating the actors as two distinct entities, as they actually are, the algorithm would erroneously amalgamate them into one unique actor.

With the NER model and the segmentation of the texts, we could not only identify senders and recipients of these correspondences but also their attributes. At this stage of the project we are particularly interested in two of them: occupation (*OCC*) and affiliations (*ORG*, *MILIT*, and *RELIG*), especially the first. Fig. 8 and Table 1 show the frequency (top 100 and top 10, respectively) with which occupations and organisations appear in the records studied, from 1642 to 1822.

5. Correspondence networks in the Portuguese empire

In this section we present the construction of the networks and their selected descriptive statistics using the current version of our dataset. We build networks based on the correspondence exchanged between the Lisbon administration and its Atlantic overseas colonies under the reign of seven monarchs, from 1642 to 1822, as shown in Fig. 3. These are directed and weighted networks, such that edges point from the sender to the recipient and weights indicate the number of documents sent. Table 2 presents a summary of the basic characteristics of these networks. We can make three observations. Firstly, as the time progresses the networks get larger in terms of the number of actors, which is related

[8] It is worth to note that one of the biggest challenges was to find the aforementioned patterns in the structure of the correspondence exchanged. Each type of correspondence has its structure, which made it difficult to find a common ground to the analysis of the whole corpus. A direct improvement from this observation, for future work, is to create specific algorithms for each type of correspondence. The corpus would have to be heavily fragmented, but this could substantially increase the accuracy of the digital text analysis.

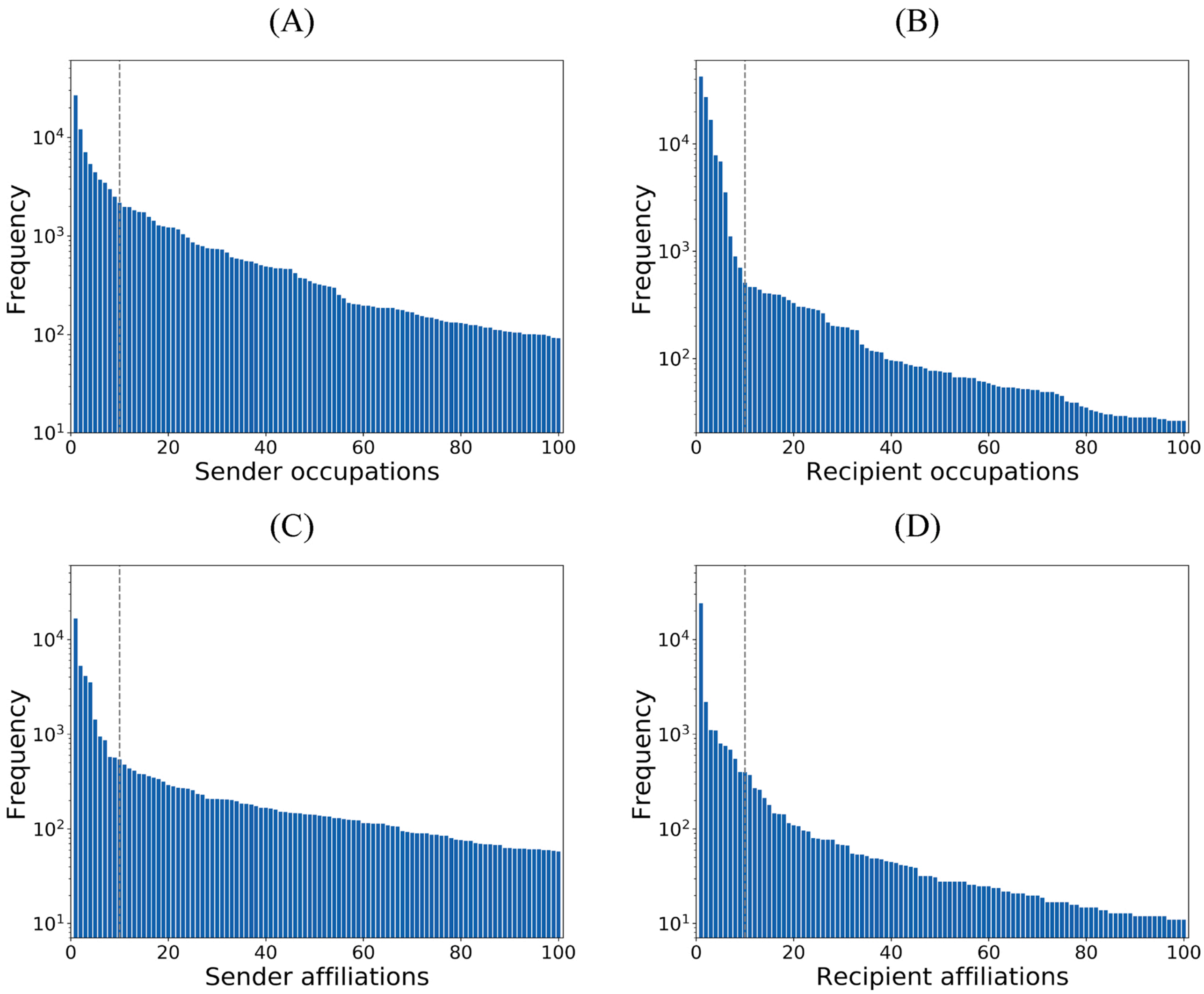


Fig. 8. Frequency with which the 100 most frequent occupations of senders (A) and recipients (B), and the 100 most frequent affiliations of senders (C) and recipients (D) appear in the letters. The dashed gray line in each panel marks the top 10 entries listed with names in Table 1.

Table 1
Top 10 occupation and affiliation for senders and recipients, as shown in Fig. 8.

Senders		Recipients	
Occupation	Affiliation	Occupation	Affiliation
Governor	Overseas Council	King	Sec. Navy Overseas Terr.
General Captain	Royal Depart. of Economy	Secretary of State	Overseas Council
Captain	Sec. Navy Overseas Terr.	Prince Regent	Royal Depart. of Economy
Secretary of State	Municipal Chamber	Queen	Foreign affairs and War
Provider	Customs	Governor	Sec. King. Royal Mercy
Vice-Rei	Infantry	General Captain	Navy Overseas Terr.
Prosecutor	Governmental Junta	Captain	Navy and Overseas Affairs
Judge (Desembargador)	National Mint	Master	Municipal Chamber
Judge (Ouvidor Geral)	Foreign affairs and War	Minister	Court
Major Captain	Secretary of the Gold	Judge (Ouvidor Geral)	Royal Affairs

Table 2
Summary of basic characteristics of the correspondence network, from 1642 to 1822. LCC stands for largest connected component. LCC, average degree and density are calculated considering both in- and out-edges. The average degree remains remarkably stable and the size of the LCC shows that the networks were very well connected albeit sparse.

Monarch	Size	Size of LCC	Average degree	Density
John IV (1640–1656)	301	265 (88 %)	2.06	0.0068
Afonso VI (1656–1683)	593	552 (93 %)	2.26	0.0034
Peter II (1683–1706)	611	542 (89 %)	2.14	0.0035
John V (1706–1750)	8279	8150 (98 %)	2.36	0.0003
Joseph I (1750–1777)	7869	7734 (98 %)	2.57	0.0003
Maria I (1777–1816)	18,506	18,341 (99 %)	2.55	0.0001
John VI (1816–1826)	2781	2728 (98 %)	2.31	0.0008

to the number of correspondence exchanged as seen in Fig. 3. Secondly, despite the different sizes, the average number of connections (total degree, i.e. in-degree plus out-degree) is remarkably stable with values in the range 2–2.5. This might be a consequence of the fact that the majority of the correspondence involves a citizen and the monarch or a government official who acts as a local hub. Thirdly, all the networks are very well connected. The percentage of actors in the largest connected component (LCC) is very high and similar for all monarchs. The reason for such a pattern might be the fact that the communication is organized around officials of the empire – the network primarily consists of people who communicated with the central actors and institutions of the empire.

The average degrees, as shown in Table 2, tell us little about the structure of the networks, especially in the case of hierarchical correspondence networks involving important actors such as kings and queens. Hence, we look at the degree distribution of these networks (Fig. 9). As expected, due to the hierarchy, the degree distributions are heavy-tailed, for both in- and out-degrees (counting persons) and weighted in- and out-degrees (counting documents), highlighting the importance and the strong presence of the monarch.

Yet, we can see the striking difference, for instance, between the in-degree distribution of John V and his successor Joseph I. The former is concentrating most of the official correspondence, while the latter, although still being the most frequent recipient, shares the stage with other important actors: secretaries of state, colonial governors, and other officials based in the colonies. The abrupt change from one monarch to its successor – and the continuity of this new pattern in the following reigns – is an indicator of the rising prominence of colonial actors (individuals and institutions). It might pose the grounds for the emergence of the multi-continental monarchy, as the new developing theory in the Luso-Brazilian historiography is proposing. We intend to

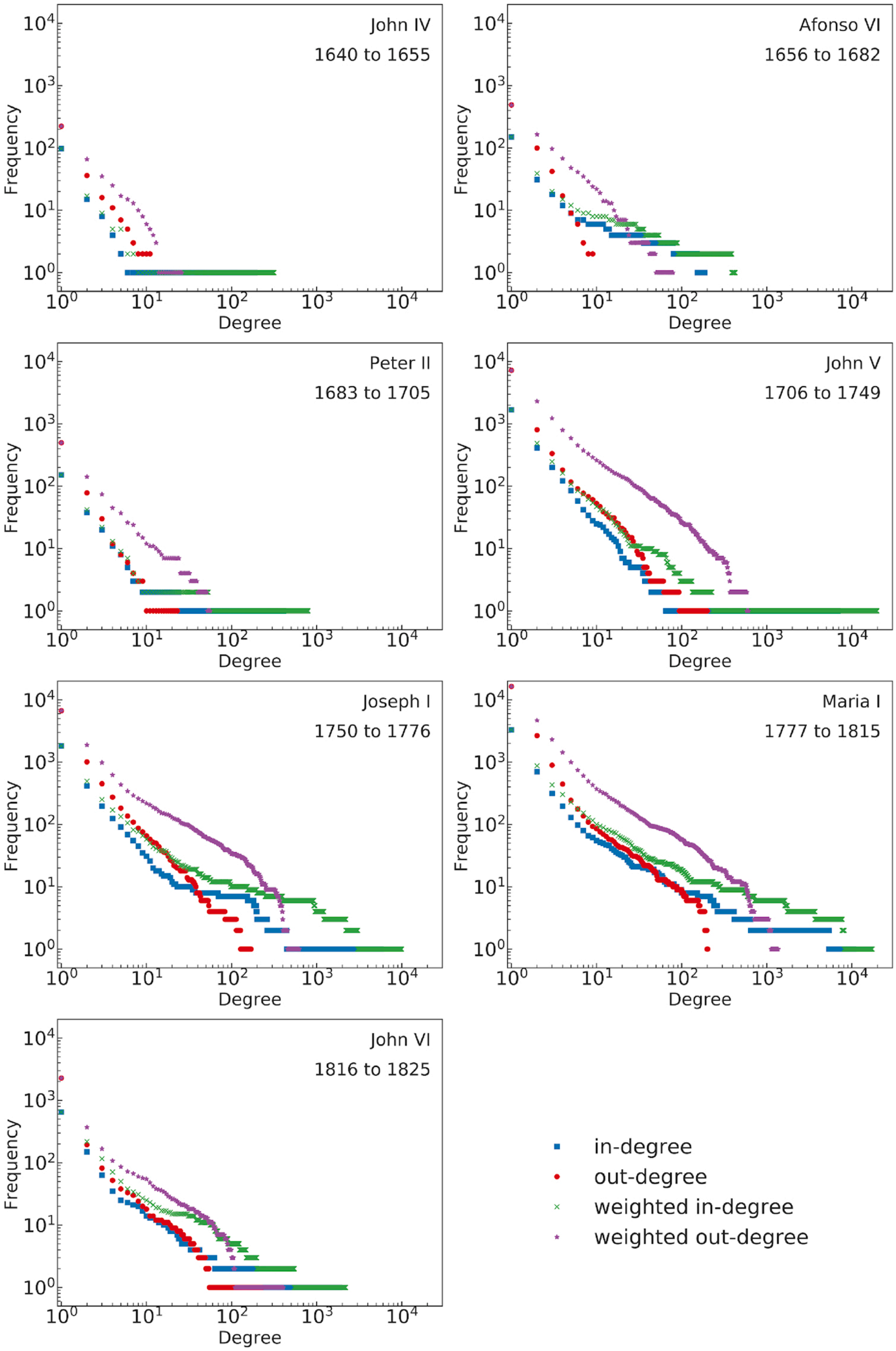


Fig. 9. Complementary cumulative frequency for degrees and weighted degrees of the correspondence networks during the reign of seven Portuguese monarchs, considering documents exchanged between 1642 and 1822. The degree distributions corroborate the idea that citizens from the colonies could reach the monarch directly, given the extremely heavy-tailed in-degree distributions. On the other hand, it is interesting to note a change in these in-degree distributions from John V to Joseph I. During the reign of the latter, more governmental officials (especially secretaries and governors) appear as important recipients of the administrative correspondence exchange between the colonies and the metropole.

further investigate how this structural change in the networks reflect changes in the political and administrative decision-making processes of the empire. We believe that a more detailed network analysis, in conjunction with a qualitative approach, can shed some light in questions like: to what extent did the Portuguese Empire become decentralised? In which sovereign matters did the officials in the colonies play an active role? How did the prominent local institutions act in defending the interests of the colony against the metropole?

Following this reasoning, we present the networks visualisations for the reigns of John V and Joseph I (Fig. 10). Apart from the monarch, the most active actors were the colonial governors and officials (see Table 3). We can see the appearance of new hubs under the reign of Joseph I, as already hinted by the degree distributions of Fig. 9. It is worth noting that, although the governors do not appear so strongly active individually, they show an important collective presence (see Fig. 8 and Table 1). Governors tended to regulate the information flow and were able to control who had access to the metropole (Fragoso et al., 2017). However, we also see that a significant number of different colonial residents could directly contact or petition to Lisbon (red nodes in the networks of Fig. 10 represent the monarchs and their respective ego-networks). On the one hand, this raises questions regarding the actual role of the governors as gatekeepers, and the decentralisation of power and the level of autonomy of local authorities. On the other hand, could the number of direct correspondents with the monarch indicate that individuals, and perhaps community leaders, had their voices heard in the metropole? Could the emergence of the multi-monarchy be a process involving not only the governmental officials but also the colonial society?

The purpose of showing the networks and descriptives in this section is primarily to give a taste of the data and the potential of the available information in answering some of the research questions that motivated this project. More detailed and substantively-oriented analyzes are the subject of the next phases of our research.

6. Discussion

Our primary motivation for writing this article was to demonstrate how we turned a massive unstructured archival content, available digitally, into network data which can be interesting to historians and (social) network scientists. We close the article with two sets of conclusions. First, technical ones related to our experiences of processing unstructured material into structured data amenable for social network analysis. Second, more substantive ones related to the advantages and disadvantages of the presented dataset for the historical research on early-modern Portuguese Empire.

On the technical side, constructing a dataset from textual information is not an easy task even if human effort is augmented with machine learning algorithms and fast computers. If we assume that manual coding a single document from our corpus would take 15 min then a single person working eight hours a day, seven days a week, would spend about 14 years coding the entire corpus. A computer trained on a small subset of human-coded documents can do that in a matter of minutes. Albeit perhaps not as accurately. One of the greatest concerns of historians regarding the use of complex automatic document processing mechanisms to perform the historical analysis is the risk of simplifying registered information, which may affect the reliability of the final research results. Therefore, we decided to use regular expressions, thanks to which we checked whether the sequences matched our specific patterns, and then, by using named entity recognition, we

Table 3
Top 10 actors (occupation and affiliation) with the highest in-degree during the reigns of John V and Joseph I, excluding the monarch.

John V	In-degree	Joseph I	In-degree
Manuel Caetano Lopes de Lavre (Secretary of the Overseas Council)	64	Francisco Xavier de Mendonça Furtado (Secretary of State, Navy and Overseas)	443
Gomes Freire de Andrade (Governor and Captain General)	43	Tomé Joaquim da Costa Corte Real (Secretary of State, Navy and Overseas)	198
Diogo de Mendonça Corte Real (Secretary of State, Navy and Overseas)	36	Martinho de Melo e Castro (Secretary of State, Navy and Overseas)	194
André Lopes de Lavre (Secretary of the Overseas Council)	34	Sebastião José de Carvalho e Melo (Secretary of the Kingdom and Royal Mercy)	181
António Guedes Pereira (Secretary of State, Navy and Conquered Territories)	24	Diogo de Mendonça Corte Real (Filho) (Secretary of State, Navy and Overseas)	154
Rodrigo César de Meneses (Governor and Captain General)	21	Gomes Freire de Andrade (Governor and General Captain)	69
Vasco Fernandes César de Meneses (Viceroy and Captain General)	18	Baltasar Manuel Pereira do Lago (Governor and General Captain)	37
Antônio Luís de Távora (Governor and Captain General)	17	Joaquim Miguel Lopes de Lavre (Secretary of Overseas Council)	34
João Jacques de Magalhães (Governor and Captain General)	17	Francisco Xavier de Mendonça (Secretary of State, Navy and Overseas)	19
Paulo Caetano de Albuquerque (Governor and Captain General)	16	Henrique Guilhon (Judge, Royal Department of Economy)	19

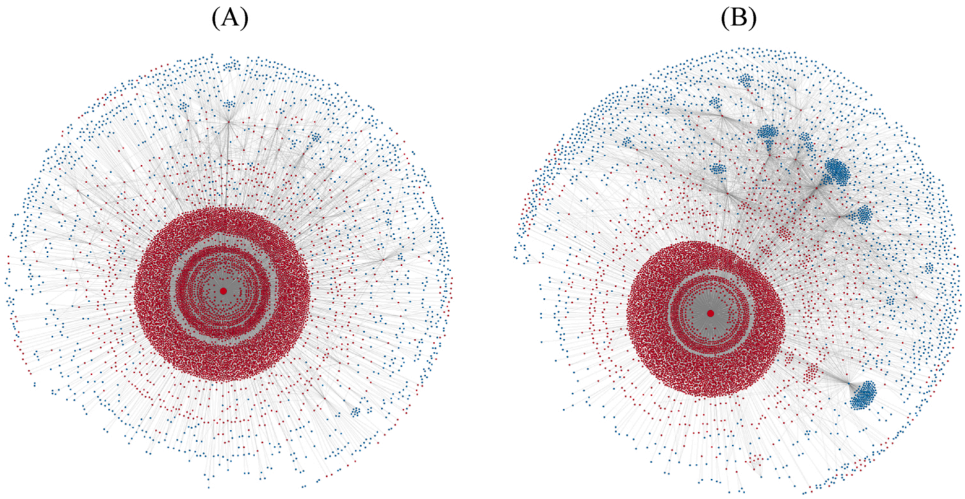


Fig. 10. Correspondence networks during the reign of (A) John V, and (B) Joseph I. Red nodes are the monarch (in the centre of the red arc) and its respective ego-network, showing that citizens of the whole empire could reach the monarch directly. Other high degree nodes correspond to government officials acting like hubs. These are mostly secretaries and governors (c.f. Table 3) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

managed to classify entities into predefined categories. It is worth noting that when preparing historical documents for further data analysis, we shall always check manually some of the stored data in such a way that the training model can learn from data and perform in a more accurate way. The more accurate the training data are, the more accurate the results of the final text classification are.

We formulate the following recommendations regarding the data challenges in extracting data from historical texts from a broader perspective. Before analyzing the text and preparing the appropriate algorithms, one should consider which words actually need such algorithms and which ones are identifiable via more straightforward keyword search. In our dataset we needed algorithms primarily to identify people (due to the specificity of Portuguese names) but also to identify and classify institutions according to their political, religious, or military character, because the vocabulary changed over time due to the administrative development of the Portuguese empire. Issues related to religious and ethnic minorities do not require separate algorithms and can be easily identified through keywords.

Regarding text structure, we consider regular expressions an extremely helpful tool for finding patterns in the text and to identify the pieces of text containing key information. We have learned that even unstructured documents can have patterns, so it might make all the difference spending some time trying to find them. In historical network analysis, we need to deal with an extremely large number of unstructured text at once and it can make the data analysis ever more complicated. Therefore, we recommend the researchers to split the study into several smaller data sets to be addressed one at a time. A small data set can be manually checked and then used as input to the machine learning models, which are in turn used on the next set of documents to be studied. We believe this can reduce the total amount of manual work and improve the quality of the models. As far as manual work is concerned, it is crucial to be guided by quality not quantity. The long hours of manual work on coding and checking the accuracy of the algorithms is by far the most "expensive" resource of a researcher. Low quality output will most likely lead to repeating other manual tasks.

The methods we describe in this article should be helpful for documents that are official correspondence or any source of plain text. The algorithms helped us to determine information about senders and recipients as well as their social attributes and geographic locations. We believe our methodology may help in setting the first steps to convert other types of historical documents into network data. As such, our approach is agnostic to the language of the documents. What needs adjusting are the regular expressions and the NER model we described in section 4.1. The model, in particular, was trained by us from scratch because the publicly available instance of the model trained on modern texts and Portuguese Wikipedia did not perform well. However, historians are working on various sources from different historical eras and various archives. It may well be that the training step will be unnecessary when used on documents describing, for instance, modern events.

On the substantive side, the corpus consists primarily of letters and other documents that can be interpreted as official but interpersonal communication: senders and recipients are persons, not only collectives or institutions. Still, the data should not be treated as a complete communication network. Letters and documents usually stay by the recipient and our corpus comes from an archive in Lisbon. As a result the majority of the documents in the corpus are sent to the officials residing in Lisbon and we are lacking the replies and other documents sent from the officials residing in Lisbon to, say, governors of overseas colonies. However, we believe that the missing replies are not significant in volume. Letters can be answered, but in general, they are not, especially for petitions, decrees, and the likes of it. The replies to the petitions might be also a decree or any other royal decision. Nevertheless, the data provides a broader perspective of what types of social and political actors, with underprivileged groups in particular, communicated with the king and the empire's capital.

We believe the dataset we are constructing has a much larger research potential than we were able to even glimpse in this article. Among the immediate research problems are for example whether the appearance of high degree actors during the reigns of subsequent monarchs (as shown in Figs. 9 and 10) is an evidence of changes in governance towards the monarch delegating some tasks to his subordinate officials. Are there differences between networks in different colonies? Do different governmental positions correspond to structurally equivalent network positions? How does the inter-institutional network look like and how does it change over time? These questions are related to the development of the multi-continental monarch theory, to which we aim to contribute.

According to the research interests that one might have, "trimming" the network might also be an interesting approach. This trimming could go in two ways. On one hand, if one is interested in, for example, studying the communication between government officials only, then one would consider removing all the low-degree nodes corresponding to other types of actors. On the other hand, if one is interested in, say, communication relations of the marginalized ethnic groups, then one might exclude those government officials who never received a document from a member of such marginalized ethnic groups. Besides, the places from where the letters were sent enable identifying the sender's localization, so one can track individuals and even determine the probability of their encounters in a particular place.

Further enrichment of the constructed dataset enabling addressing other interesting historical research questions is on our agenda. We mention two of them here. Firstly, the data provides not only the information about who sends a document to whom, but also about what. We plan to use methods of *topic modeling*, such as Latent Dirichlet Allocation (Blei et al., 2003), to cluster the documents to groups according to their content. It might be possible to identify main topics and their changing popularity throughout the whole period and across geographical locations. Secondly, the authors write the documents mentioning other people, events and places. We are investigating methods through which our corpus can be used as a source of such "indirect" information about events, persons, geographical locations and social relations between them. Similar goals were pursued by Warren et al. (2016) in their analysis of the Oxford Dictionary of National Biography. In other words, to construct a network data that also involves people who were not senders or recipients, but were mentioned by others in their communications.

Acknowledgements

Research presented in this article is a part of the Mapping the Atlantic Portuguese Empire project (https://mape.netlify.app). Authors are very grateful to two anonymous reviewers for the work they invested in giving constructive comments that helped us to significantly improve the article. Agata Błoch thanks (Polish) National Science Centre (grant 2017/27/N/HS3/01104), and the Tadeusz Manteuffel Institute of History of Polish Academy of Sciences for support. Authors also thank Jessica Joia, Bartosz Lemiec, and Monika Pawluczuk for their contributions to data preparation and processing. We are also grateful to Lino Oliveira Lopes-De Matos and his students from University Marie Curie-Sklodowska in Lublin for their help in annotating the source documents.

References

Adams, J., 2019. Gathering Social Network Data. Sage Publications.

Ahnert, R., Ahnert, S.E., 2015. Protestant letter networks in the reign of Mary I: a quantitative approach. ELH 82 (1), 1–1.

Alexander, M.C., Danowski, J.A., 1990. Analysis of an ancient network: personal communication and the study of social structure in a past society. Soc. Networks 12 (4), 313–335.

Baker, W.E., Faulkner, R.R., 1993. The social organization of conspiracy: illegal networks in the heavy electrical equipment industry. Am. Sociol. Rev. 837–860.

Bellotto, H.L., 2004. Arquivos Permanentes: Tratamento Documental. FGV editora.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3 (Jan), 993–1022.

Boletim do Arquivo Histórico Colonial, 1950. Arquivo Histórico Colonial, Vol. 1.

Booth, A., Connell, S.E., Coolahan, M.-L., Boggs, J., Flanders, J., Kelly, D., Mapp, R., Martin, W., 2017. Only Connect!: Intertextuality, Circulation, and Networks in Digital Resources for Women's Writing. *DH*.

Boschi, C.C., 2011. O Brasil-Colônia nos arquivos históricos de Portugal: Roteiro sumário. Alameda.

Bourke, E., 2017. Female involvement, membership, and centrality: a social network analysis of the Hartlib Circle. Lit. Compass 14 (4), e12388.

Breure, A.S., Heiberger, R.H., 2019. Reconstructing science networks from the past. J. Historical Network Res. 3, 92–117.

Carley, K.M., 1997. Network text analysis: the network position of concepts. Text Analysis Social Sci. Meth. Drawing Statistical Inferences Texts Transcripts 4, 79–100.

Carley, K.M., Palmquist, M., 1992. Extracting, representing, and analyzing mental models. Soc. Forces 70 (3), 601–636.

Cline, D.H., 2020. Athens as a small world. J. Historical Network Res. 4, 36–56.

Coolahan, M.-L., 2017. The material cultures of early modern women's writing. Early Mod. Women 11 (2), 151–154.

Curran, B., Higham, K., Ortiz, E., Vasques Filho, D., 2018. Look who's talking: two-mode networks as representations of a topic model of New Zealand parliamentary speeches. PLoS One 13 (6), e0199072.

Da Costa, E.V., 1997. Da senzala à colônia. Editora UNESP.

Da Costa, E.V., 2000. The Brazilian Empire: Myths and Histories. University of North Carolina Press.

Danowski, J.A., 2009. Inferences from word networks in messages. The Content Analysis Reader 421–429.

Diesner, J., Carley, K.M., 2005. Revealing social structure from texts: Meta-matrix text analysis as a novel method for network text analysis. Causal Mapping for Research in Information Technology. IGI Global, pp. 81–108.

Diesner, J., Carley, K.M., Tambayong, L., 2012. Extracting socio-cultural networks of the Sudan from open-source, large-scale text data. Comput. Math. Organ. Theory 18 (3), 328–339.

Edelstein, D., Kassabova, B., 2018. How England fell off the map of Voltaire's enlightenment. Mod. Intellect. Hist. 1–25.

Edelstein, D., Findlen, P., Ceserani, G., Winterer, C., Coleman, N., 2017. Historical research in a digital age: reflections from the mapping the republic of letters project. Am. Hist. Rev. 122 (2), 400–424.

Fernandes, F., 1969. Beyond Poverty: The negro and the mulatto in Brazil. Journal de La Société Des Américanistes 58, 121–137.

Fragoso, J., Gouvêa, M.de F., 2009. Monarquia pluricontinental e repúblicas: Algumas reflexões sobre a América lusa nos séculos XVI-XVIII. Revista Tempo 14 (27), 49–63.

Fragoso, J.L.R., Jucá de Sampaio, A.C., 2012. Monarquia Pluricontinental e a governança da terra no ultramar atlântico luso. Séculos XVI-XVIII. Mauad.

Fragoso, J.L.R., Monteiro, N.G., Costa, A., 2017. Um reino e suas repúblicas no Atlântico: Comunicações políticas entre Portugal. Brasil e Angola nos séculos XVII e XVIII. Civilização Brasileira.

Franzosi, R., 1997. Mobilization and counter-mobilization processes: from the "red years"(1919–20) to the "black years"(1921–22) in Italy. Theory Soc. 26 (2), 275–304.

Franzosi, R., 1998. Narrative as data: linguistic and statistical tools for the quantitative study of historical events. Int. Rev. Soc. Hist. 43 (S6), 81–104.

Franzosi, R., 2004. Content analysis. Handbook of Data Analysis 547, 566.

Franzosi, R., 2008. Content analysis: objective, systematic, and quantitative description of content. Content Anal. 1 (1), 21–49.

Fuhse, J., Stuhler, O., Riebling, J., Martin, J.L., 2019. Relating social and symbolic relations in quantitative text analysis. A Study of Parliamentary Discourse in the Weimar Republic. Poetics.

Germerodt, F., 2020. Networking in the early roman empire: pliny the younger. J. Historical Network Res. 4, 252–270.

Gilles, G., 2020. Family or faction? The political, social and familial networks discerned from Cicero's letters during the civil war between Caesar and pompey. J. Historical Network Res. 4, 114–155.

Gondal, N., McLean, P.D., 2013a. Linking tie-meaning with network structure: variable connotations of personal lending in a multiple-network ecology. Poetics 41 (2), 122–150.

Gondal, N., McLean, P.D., 2013b. What makes a network go round? Exploring the structure of a strong component with Exponential Random Graph Models. Soc. Networks 35 (4), 499–513.

Grandjean, M., 2016. Archives Distant Reading: Mapping the Activity of the League of Nations' Intellectual Cooperation. *Undefined*. Digital Humanities. Kraków, Poland.

Grandjean, M., 2017. Multimode and Multilevel: Vertical Dimension in Historical and Literary Networks. *DH*.

Guedes, R., 2013. Dinâmica imperial no Antigo Regime português. Mauad.

Herlihy, D., Klapisch-Zuber, C., 1985. Tuscans and their families: A study of the Florentine Catasto of 1427. Yale University Press, New Haven.

Herlihy, D., Klapisch-Zuber, C., Litchfield, R.B., Molho, A., 2002. Online Catasto of 1427. Machine readable data file based on D. Herlihy and C. Klapisch-Zuber. Census and Property Survey of Florentine Domains in the Province of Tuscany, 1427-1480.

Jimerson, R.C., 2003. Archives and memory. OCLC Systems & Services: International Digital Library Perspectives.

Kenna, R., Mac Carron, P., 2016. Maths meets myths: network investigations of ancient narratives. J. Phys. Conf. Ser. 681, 012002.

Kent, D.V., 1978. The Rise of the Medici: Faction in Florence. Oxford University Press, USA, pp. 1426–1434.

Köstner, E., 2019. Trimalchio's last will. J. Historical Network Res. 3, 1–29.

Köstner, E., 2020. Genesis and Collapse of a Network. The Rise and Fall of Lucius Aelius Seianus. J. Historical Network Res. 4, 225–251.

Laumann, E.O., Marsden, P.V., Prensky, D., 1989. The boundary specification problem in network analysis. Res. Methods Social Network Anal. 61, 87.

Lee, M., Martin, J.L., 2015. Coding, counting and cultural cartography. Am. J. Cult. Sociol. 3 (1), 1–33.

Mac Carron, P., Kenna, R., 2012. Universal properties of mythological networks. EPL (Europhysics Letters) 99 (2), 28002.

Mac Carron, P., Kenna, R., 2013. Network analysis of the íslendinga sögur–the sagas of icelanders. Eur. Phys. J. B 86 (10), 407.

Magnini, B., Negri, M., Prevete, R., Tanev, H., 2002.). A WordNet-based approach to named entites recognition. COLING-02: SEMANET: Building and Using Semantic Networks.

McGee, F., During, M., Ghoniem, M., 2016. Towards Visual Analytics of Multilayer Graphs for Digital Cultural Heritage.

McLean, P.D., Gondal, N., 2014. The circulation of interpersonal credit in Renaissance Florence. Eur. J. of Sociol. Archives Européennes de Sociologie 55 (2), 135–176.

McShane, B.A., 2018. Visualising the reception and circulation of early modern nuns' letters. J. Historical Network Res. 2 (1), 1–25.

Mohr, J.W., 1994. Soldiers, mothers, tramps and others: discourse roles in the 1907 New York City charity directory. Poetics 22 (4), 327–357.

Mohr, J.W., Duquenne, V., 1997. The duality of culture and practice: poverty relief in New York City, 1888-1917. Theory Soc. 26, 305–356.

Moretti, F., 2013. Distant Reading. Verso Books.

Novais, F.A., 1995. Portugal e Brasil na crise do antigo sistema colonial (1777-1808).

Padgett, J.F., Ansell, C.K., 1993. Robust action and the rise of the medici, 1400-1434. Am. J. Sociol. 98 (6), 1259–1319.

Padgett, J., Prajda, K., Rohr, B., Schoots, J., 2019. Political Discussion and Debate in Narrative Time: the Florentine Consulte E Pratiche, 1376–1378. *Poetics, Special Issue on 'Discourse, Meaning, and Networks: Advances in Socio-Semantic Analysis*.

Popping, R., 2000. Computer-assisted Text Analysis. Sage.

Popping, R., 2003. Knowledge graphs and network text analysis. Soc. Sci. Inf. 42 (1), 91–106.

Popping, R., Roberts, C.W., 1997. Network approaches in text analysis. Classification and Knowledge Organization. Springer, pp. 381–389.

Prado, C.P., 1957. Formação do Brasil contemporâneo, Vol. 1. Editora Brasiliense.

Prado, C.P., 1967. The Colonial Background of Modern Brazil. University of California Press.

Ratcliff, J.W., Metzener, D.E., 1988. Pattern-matching-the gestalt approach. Dr Dobbs Journal 13 (7), 46.

Riva, G.F., 2019. Network analysis of medieval manuscript transmission. J. Historical Network Res. 3, 30–49.

Roorda, D., Bos, E.-J., van den Heuvel, C., 2010. Letters, Ideas and Information Technology: Using Digital Corpora of Letters to Disclose the Circulation of Knowledge in the 17th Century. *DH*, pp. 211–213.

Rosillo-López, C., 2020. Informal political communication and network theory in the late roman republic. J. Historical Network Res. 4, 90–113.

Sibille, C., 2011. LONSEA – der Völkerbund in neuer sicht. Eine Netzwerkanalyse Zur Geschichte internationaler organisationen. Studies Contemporary History 8, 475–483.

Van Den Heuvel, C., 2015. Mapping knowledge exchange in early modern Europe: intellectual and technological geographies and network representations. Int. J. Humanit. Arts Comput. 9 (1), 95–114.

Van Den Heuvel, C., Weingart, S.B., Spelt, N., Nellen, H., 2016. Circles of confidence in correspondence: modeling confidentiality and secrecy in knowledge exchange networks of letters and drawings in the early modern period. Nuncius 31 (1), 78–106.

Vozár, Z., 2018. Metadata for the Middle Ages: A Network Analysis of Manuscriptorium.com.

Warren, C.N., Shore, D., Otis, J., Wang, L., Finegold, M., Shalizi, C., 2016a. Six degrees of Francis Bacon: a statistical method for reconstructing large historical social networks. DHQ: Digital Humanities Quarterly 10 (3).

Warren, C.N., Shore, D., Otis, J., Wang, L., Finegold, M., Shalizi, C., 2016b. Six degrees of Francis Bacon: a statistical method for reconstructing large historical social networks. DHQ: Digital Humanities Quarterly 10 (3).

Wieneke, L., Düring, M., Silaume, G., Lallemand, C., Croce, V., Lazzarro, M., Nucci, F., Pasini, C., Fraternali, P., Tagliasacchi, M., 2013. Building the social graph of the history of European integration. International Conference on Social Informatics 86–99.

Wieneke, L., During, M., Silaume, G., Lallemand, C., Croce, V., Lazzarro, M., Nucci, F., Pasini, C., Fraternali, P., Tagliasacchi, M., 2014. HistoGraph–A visualization tool for collaborative analysis of historical social networks from multimedia collections. Proceedings of 18th International Conference Information Visualisation (IV), 2014 Conference.

Xavier, Â.B., Silva, C.N.da., 2016. O governo dos outros: Poder e diferença no império português. Imprensa de Ciências Sociais.

Xavier, Â.B., Palomo del Barrio, F., Stumpf, R., 2018. Monarquias ibéricas em perspectiva comparada (séculos XVI-XVIII): Dinâmicas imperiais e circulação de modelos políticos-administrativos. Imprensa de Ciências Sociais.