université PARIS-SACLAY

Si la thèse a été préparée dans le cadre d'une **cotutelle internationale de thèse**, logo de l'université étrangère co-délivrant le doctorat **(sinon, à enlever)**

# Analyse Visuelle pour l'Analyse de Réseaux Sociaux Historiques

*Visual Analytics for Historical Network Research*

**Thèse de doctorat de l'université Paris-Saclay et de Telecom Paris**

École doctorale n°580 : Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat: Informatique
Graduate School : Informatique et Sciences du Numérique
Référent : Faculté des sciences d'Orsay

Thèse préparée au Laboratoire interdisciplinaire des sciences du numérique (Université Paris-Saclay, CNRS, Inria), et à Telecom Paris, sous la direction de Jean-Daniel FEKETE, Directeur de recherche et la co-direction de Christophe Prieur, Professeur des universités.

**Thèse soutenue à Paris-Saclay, le JJ mois AAAA, par**

**Alexis PISTER**

**Composition du jury**

| | |
|---|---|
| **Prénom Nom**<br>Titre, Affiliation | Président ou Présidente |
| **Prénom Nom**<br>Titre, Affiliation | Rapporteur & Examinateur / trice |
| **Prénom Nom**<br>Titre, Affiliation | Rapporteur & Examinateur / trice |
| **Prénom Nom**<br>Titre, Affiliation | Examinateur ou Examinatrice |
| **Prénom Nom**<br>Titre, Affiliation | Examinateur ou Examinatrice |
| **Prénom Nom**<br>Titre, Affiliation | Directeur ou Directrice de thèse |

THESE DE DOCTORAT

NNT : 2020UPASA001

université PARIS-SACLAY

ÉCOLE DOCTORALE
Physique et ingénierie : Electrons, Photons, Sciences du vivant (EOBE)

**Titre:** titre (en français).....
**Mots clés:** 3 à 6 mots clefs (version en français)

**Résumé:**Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

**Title:** titre (en anglais).....
**Keywords:** 3 à 6 mots clefs (version en anglais)

**Abstract:** Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

# Contents

# 2 Related Work

Social historians rely on textual historical documents to draw socio-economic conclusions about the past. They read and analyze the documents they can find from a period and subject of interest, and make their conclusions after analyzing them and cross-referencing the information they found. Several methods have been developed in History to extract and analyze the information contained in the documents in a rigorous way, such as qualitative analysis, quantitative methods, or HSNA. HSNA is a method coming from Sociology consisting in modeling the relational information mentioned in the documents—such as family, business, or friendship ties—in a network, to be able to characterize and explain social behaviors through the description of the network's structure [**?**, **?**]. HSNA is directly inspired by SNA, which is a well-known method in sociology that sociologist theorized to understand and describe real world social relationships modeled as networks [**?**, **?**]. Historians appropriated this methods, by extracting relationships from historical documents. The specificy of HSNA is therefore the modeling of the network from the historical documents—which are at the core of the historical work [**?**]—and the integration of the time aspect which is often disregarded in traditional SNA. Historians typically use social network visualization tools to confirm or generate new hypotheses once they successfully constructed their network [**?**]. In this chapter, we therefore present a general overview of the fields of SNA (**??**), HSNA (§2.2), and Social Network Visualization (§2.4).

## 2.1 Visualization

Visualization is often defined as "the use of computer-supported, interactive, visual representations of data to amplify cognition"" [**?**]. Graphically displaying data allows us to leverage our visual system to gain a better acquisition of knowledge, leading to better decision-making, communication, and potential discoveries. The field of visualization can be split in three sub domains: **Scientific visualization** focus on visualizing continuous physically based data such as weather, astrophysics, and anatomical data, sometimes produced with simulations whereas **Information Visualization** is centered around the visualization of discrete abstract data points, often multidimensional. **Visual Analytics** emerged later from Information Visualization by mixing data mining and more complex analysis process with traditional information visualization displays. We focus in this thesis on the two former branches of visualization, as social scientists use both information visualization and visual analytics systems to gain insight on the structure of the networks they are studying.

### 2.1.1 Information Visualization

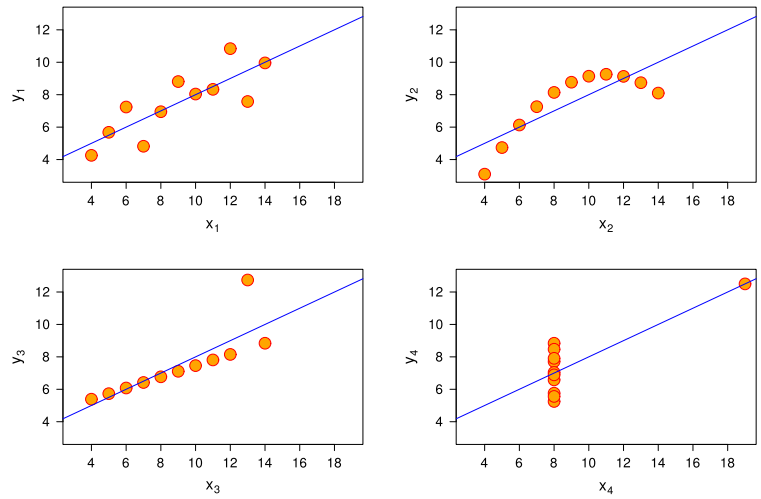


Figure 2.1 – Anscombe quartet. The four datasets have the same descriptive statistics (average, variance, correlation coefficient) but very different structures. Image from [**?**].

Information Visualization focus on displaying abstract data to amplify cognition and gain insight on real world phenomena [**?**]. History is filled with classical examples of visual data displays which helped understand specific events, such as Minard's map of Napoleon's march in Russia [**?**], or Snow's dot map of cholera cases in London which showed the proximity between street pumps and cholera infections [**?**]. If several examples of information visualization can be found thorough history, it mainly developed as a scientific field in the 1960s with Tukey's work on data analysis and visualization [**?**] and Bertin's publication of Semiology of graphics [**?**]. In this foundational work, Bertin described and organized the different visual elements usable in graphical information displays, and linked them to data features and relations types. Michael Friendly writes that "To some, this appeared to do for graphics what Mendeleev had done for the organization of the chemical elements" [**?**]. The development of computer science and the rise of hardware capabilities during the same time created a big need for data visualization. The amount of data stored increased exponentially [**?**] and descriptive statistics were not enough to understand the underlying structure of the amount and diversity of produced data. Visualization, leveraging the human visual system, allowed to rapidly see the hidden structure of a dataset and detect interesting and unexpected patterns very often unseen with classical statistical methods. One classical illustration of this is Anscombe's quartet [**?**] which consists of four datasets of points in $\mathbb{R}^2$ with the same statistical measures (mean, variance, correlation coefficient, etc.) but with very different structures, that plotting the data show immediately. The four datasets are illustrated in Figure 2.1.

A large number of visualization techniques emerged to make sense of the di-

versity of data produced, such as multidimensional, temporal, spatial, or network data [?]. Instead of using taxonomies classifying graphics into categories such as histograms, pie charts, and stream graphs, some theorized how to describe graphics in a more systematic and structural way. In 1993, Wilkinson extended Bertin's work and developed the Grammar of Graphics [?] as a way to describe the deep structure unifying every possible graphics, thus allowing to characterize and create graphics using common terms and rules. In this framework, a graphic can be defined as a function of six components: data (a set of data points and attributes from a dataset), transformations (statistical operations which modify the original data, e.g., mean and rank transformations), scales (e.g., linear and log scales), coordinate systems (e.g., cartesian and polar coordinate systems), elements (graphical marks such as rectangular or circular marks, and their aesthetics, e.g., color and size), and guides (additional information such as axes and legend). Many well-known visualization toolkits are now based on this framework, such as vega and ggplot, as it allows great expressiveness and reusability for graphic creation. Visualization allows to gain insight on the structure of a given data, and has traditionally been used for confirmation and communication purposes, for example to verify hypothesis on empirical sciences, and later on to communicate findings. Visualization is also used to communicate information to wider general audiences, for example in the context of data journalism to support a point.

### 2.1.2 Visual Analytics

Visualization can also be used for exploratory aims, to gain new insights on the general structure of the data and potentially generate new hypotheses. This process has been characterized by Tukey in 1960 as *Exploratory Data Analysis* [?] and consist in trying to characterize the structure of a dataset with the help of visualization and statistical measurements. Visual exploration is enhanced by direct manipulation interfaces through interaction and usually follows the information-seeking mantra formalized by Schneiderman: "Overview first, zoom and filter, then details-on-demand"" [?]. It allows users to first have a visual overview of the data and get an idea of its overall structure, to then change the point of focus to highlight interesting patterns with the help of filtering, querying, sorting, and zooming mechanisms. As the average size of datasets keeps growing, exploratory tools are often needed to make sense of large datasets and generate interesting hypotheses.

More recent visual exploration interfaces also incorporate automatic analytical tools along with graphical displays, letting users apply data mining algorithms directly in the exploratory loop. This coupling of visualization and data mining has been defined as Visual Analytics (VA) and is still undergoing lots of research. Keim and al. define it as "a combination of automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data" [?].

It is defined around the generation of knowledge using visualizations and mod-

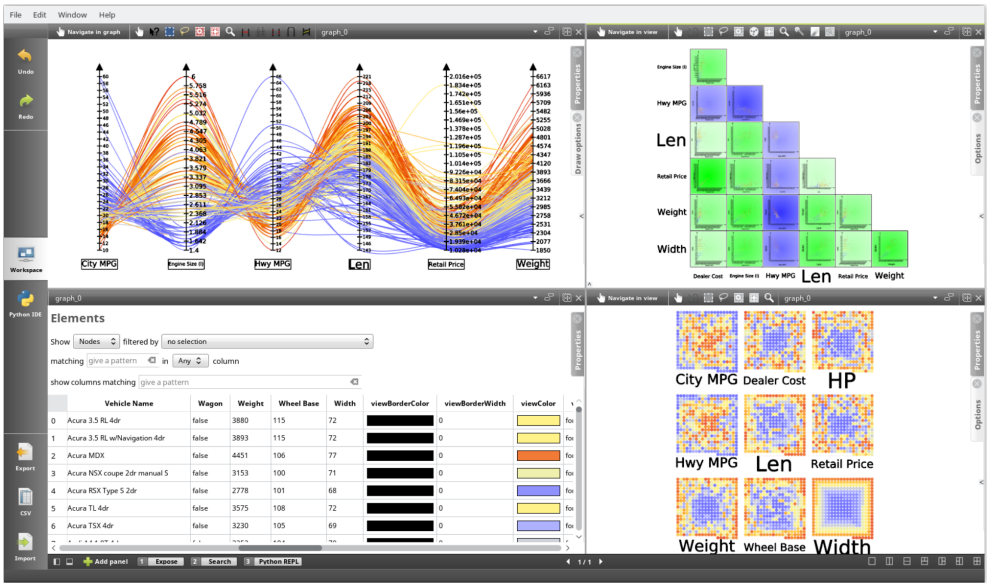


Figure 2.2 – TULIP software designed for application-independant network visual analytics [**?**]. The view shows a dataset among multiple interactive coordinated views. Users can also apply data mining algorithms on the data to extract interesting patterns.

els of the data, that the user generates and explores using interaction. VA systems have been developed in various empirical domains, such as biology, astronomy, engineering, and social sciences, as it allow to rapidly gain insight on the structure of various potentially large dataset, while generating and refuting hypotheses. Figure 2.2 shows the TULIP system, a VA system developed for the analysis of network data.

## 2.2 Quantitative Social History

Social History is a branch of history which aims at studying socio-economic aspect of past societies, with a focus on groups instead of specific individuals. Charles Tilly argues that its goal is to "(i) documenting large structural changes, (2) reconstructing the experiences of ordinary people in the course of those changes, and (3) connecting the two" [**?**]. If the purpose of social history remained the same across time, methods and formalisms have evolved since its beginning in the 1930s. Specifically, the rise of computer science led to the development of quantitative history methods in the 1960s—now often referred as Digital Humanities—which brought new ways of grounding results in numbers and quantitative models, instead of solely relying on qualitative inspection and citations of historical documents [**?**]. We discuss in this section the evolution of Social History from the context of its beginning to the use of more recent quantitative approaches.

### 2.2.1 History, Social History and Methodology

The concept of History is hard to define as its practice and codes highly evolved through time. Prost writes that "History is what historians do. The discipline called history is not an eternal essence, a Platonic idea. It is a reality that is itself historical, i.e. situated in time and space, carried out by men who call themselves historians and are recognized as such, received as history by various publics [?]." Retrospectively, History of a given time can thus be characterized by the different historical work produced at that time. Nevertheless, history can be characterized as the collection and study of historical documents to study and describe the past. As Langlois and Seignobos write, "The search for and the collection of documents is thus a part, logically the first and most important part, of the historian's craft" [?]. History emerged as a field with its own rules, conventions and journals in the 1880s from faculties of letters, to counterbalance previous history works which were judged as too "literary" [?]. A that time, two facets characterize the field, which are sometimes overlapping: one is political whereas the other one is methodological. The former aspect of history serves to create a shared story for the studied country and a sense of unity to its citizens. Antoine Prost says that "it's through history than France thinks itself" [?]. The latter aspect of history constitute a methodology to describe the past through methodic inspection of historical sources, in the aim of inferring dated facts about the past and trying to minimizing possible bias. Historical documents are thus at the core of the work of historians and having to cite historical documents and previous peers work to new claims is primordial to be considered as rigorous History work. However, methodological and epistemological facets (how historians should read and analyze their sources, how to cite them, what to report/not report etc.) of History have not been studied and discussed for a long time, until the end of the 1980s. Some historians were interested in historiography [?], but none were going to philosophical and epistemological reflexions of the History discipline. For Lucien fèbvre, philosophising was even constituting a "capital crime" [?, ?].

Retrospectively, we can still observe shifts in the objects of study of historians through time, and their relation to sources. History was at first mainly event-centered and was focusing in characterizing central figures of the past like rulers and artists or shed light on central events like wars or political crisis. This narrative approach to history has been criticized for its open interpretation of historical documents, which can introduce bias from the authors [?].

In the 1930s, March Bloch and Lucien fèbvre detached from traditional history by creating the "Annales school" (Ecole des Annales) which aimed at placing the human as a component of a broader sociological, political, and economic system with influences between each other [?]. They strongly advised to exhaustively search from archives, to ground historical results in documents, texts and numbers. This new way of studying past events and societies became successful in a profession in crisis, by bringing a new lens of study on various societal subjects more

grounded in sources and with a better intelligibility. This school of thought can be seen as one of the biggest milestones for Social History, which focuses on the socio-economical aspects of societies and their changes through time, rather than an event-centric view of History. For example, in his thesis, Ernest Labrousse—a well known figure of Social History—tries to describe and explain the economic crisis of France at the end of the "Ancien Régime"[1] through the evolution of the economic power of different social groups such as farmers, workers, property owners etcinstead of solely describing memorable facts about the period [?]. Social History continued to evolve since the 1930s, introducing new methods and concepts, but always with the goals to describe periods and historical facts through a sociological lens and with a strong focus on sources and traceability.

### 2.2.2 Quantitative History

With the development of statistical methods and Computer Science, quantitative approaches of History emerged in the 1960s with the goal of analyzing numeric data directly extracted from historical documents. Economists led this first wave of quantification by studying past events using economical concepts and data. This approach, called "new economic history" or "cliometrics" was popularized by Robert Fogel's study on the economic impact of the development of railroads in America [?] and Fogel and Engerman's controversial work on the economy of slavery [?]. In the latter study, they extracted numbers of a sample of 5000 bills of slave sales from New Orleans to support the controversial claims that slavery was economically viable and that slaves had a decent material life, which brought up heated debate among the scientific community and the mainstream audience [?]. These kinds of approaches rapidly started to be used in other related domains such as demography, social history, and political history, sometimes rebranded as "new social history" ad "new political history" [?]. As extracting the data from raw documents and uploading it in computers—which were shared among whole departments—was very time-consuming at that time, "new history" projects often relied on a high division of labor among researchers, assistants, and students who operated with punch card operators [?]. Many saw the future of social sciences in computer programming, as Le Roy-Ladurie who wrote in 1968 "The historian of tomorrow will be a programmer, or he will not exist" [?].

However, quantitative methods started to be criticized in the 1980s with a vague of disillusionment, for several reasons. Stone was the first to raise his voice in 1979, after participating himself in several of those ambitious projects: "It is just those projects that have been the most lavishly funded, the most ambitious in the assembly of vast quantities of data by armies of paid researchers, the most scientifically processed by the very latest in computer technology, the most mathematically sophisticated in presentation, which have so far turned out to be the

[1] The "Ancien Régime" is an historical period of France which starts from the beginning of the reign of the Bourbon house at 1589 until the Revolution in 1789.

most disappointing" [?]. First, many researchers of this first wave dispensed themselves of source criticism, leading to simplification, anachronisms—such as using modern analytical categories and indices like the GDP—, and taking the numeric data from historical documents as objective. These problems could be in part explained by the fact that the work process was highly divided, meaning that the people analyzing the data did not necessarily inspect and read the original historical documents in depth. Secondly, the popularity of these methods made practitioners forget about the many biases inherent to statistics, such as the sampling bias, or the fact that historical data is essentially uncomplete data. This resulted in the computation of long data series and aggregates which were sometimes nonsensical given the gaps in the sources [?]. Finally, many historians raised their voice against the study of long-term trends instead of focusing on specific events and individuals. They challenged aggregations procedures and its assumptions, trying to go back to a more complex history by pointing that phenomena have to be studied and understood through several scales [?]. Indeed, computing correlations and aggregates at a national level greatly simplify complex phenomena, and misses rare cases along specific group and individual related behaviours. Still, if their adoption remains slow and sometimes criticized among historians, quantitative methods provide tools to store, explore, and analyze historical documents systematically if used appropriately (i.e. not trying to bias the analysis, and not losing the trace of the original sources), especially that those methods highly evolved since the 1960s.

Guildi and Aritage went as far as criticizing the decrease of interest of historians working in archives [?]. Approaches using digital methods and tools are nonetheless more and more popular, sometimes more recently referred to under the umbrella term Digital Humanities (DH). If their adoption remains slow and sometimes criticized among historians, they still provide tools to store, explore, and analyze historical documents systematically if used appropriately (i.e. not trying to bias the analysis, and not losing the trace of the original sources). DH can also provide infrastructures and tools to study large historical databases which is more complicated to do by hand, as with the Venice Time Machine project [?] which aims at digitizing and analyzing thousands of documents from the archives of Venice to understand the political, geographical, and sociological dynamics of the cities across generations and centuries.

### 2.2.3 Digital Humanities

Digital Humanities is sometimes described as the second wave of computational social sciences [?]. The term has gained popularity since the 2010s and refer to "research and teaching taking place at the intersection of digital technologies and humanities. Digital Humanities aims to produce and use applications and models that make possible new kinds of teaching and research, both in the humanities and in computer science (and its allied technologies). Digital Humanities also studies the impact of these techniques on cultural heritage, memory institutions, libraries, archives and digital culture." [?]. If the first wave of computational social sciences

focused a lot on statistical methods such as regression models, correlation testing, and descriptive measures (mean, median, and variance) to make conclusions, digital humanities focuses more on the use of digital tools for exploration, teaching, and communication of humanities datasets and concepts, leveraging design, infographics and interactive systems [?]. In the context of historical research, the term Digital History have been coined as "an approach to examining and representing the past that works with the new communication technologies of the computer, the Internet network, and software systems. On one level, digital history is an open arena of scholarly production and communication, encompassing the development of new course materials and scholarly data collections. On another, it is a methodological approach framed by the hypertextual power of these technologies to make, define, query, and annotate associations in the human record of the past. To do digital history, then, is to create a framework, an ontology, through the technology for people to experience, read, and follow an argument about a historical problem." [?] Research which label itself as Digital History pivot around the curation and digitization of historical archives, the identification of historical concepts through computational and exploration methods, but also their communication to the general audience through digital technologies.

A lot of Digital History projects are thus multidisciplinary by essence and involve several teams of researchers, such as the Republic of Letters project which consisted in digitizing, storing, and exploring letters of scholars across the world, in a common hub and using shared visualization tools [?]. It resulted in the elaboration of curated datasets and visualizations concerning the correspondence of various scholars such as Voltaire, Benjamin Franklin (see Figure 2.3), or John Locke, accessible in the same place by researchers and the general audience. With modern technologies and infrastructures, it also becomes possible to study large historical databases—often labeled under the term "big data"—as with the Venice Time Machine project [?] which aims at digitizing and analyzing thousands of documents from the archives of Venice to understand the political, geographical, and sociological dynamics of the cities across generations and centuries. Yet, some historians raised concern about this type of project, fearing that it could rapidly bring the same type of issues that we saw during the first wave of quantification, especially for big projects involving many actors and high ambitious goals. Many projects which claim themselves as Digital History also leverage new methods compared to the 1960s and 1970s, such as the use of network methods and concepts [?]. Examples are the Viral Texts [?] and Living with Machines [?] projects which respectively study nineteenth-century newspapers and the industrial revolution by translating their sources into analyzable networks. We discuss more in depth the related work of network analysis for historical research in **??**.

## 2.3 Historical Social Network Analysis

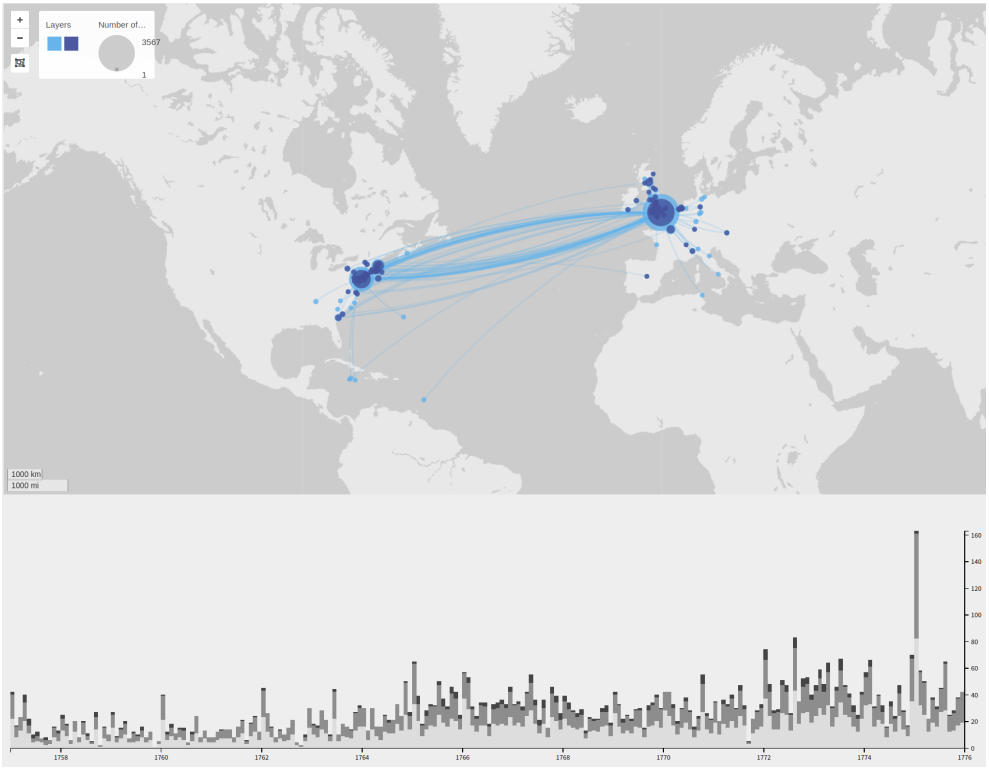


Figure 2.3 – Correspondence letters of Benjamin Franklin and his close relationships, using a map and an histogram, accessible online on the republic of letter website [**?**].

Historians started to use network analysis to study relational structures and phenomena of past societies in the 1980s, using similar methods developed by sociologist under the label of SNA. SNA is defined as an "approach grounded in the intuitive notion that the patterning of social ties in which actors are embedded has important consequences for those actors. Network analysts, then, seek to uncover various kinds of patterns. And they try to determine the conditions under which those patterns arise and to discover their consequences" [?]. the use of networks emerged in response to traditional sociology methods using pre-defined taxonomies and social categories to understand and explain sociological behaviors and phenomena, which could introduce bias. By modeling real observed social relationships and interactions with networks and by using mathematical and statistical methods to study those, sociologists have been able to explain sociological phenomena and describe sociological interactions through their direct observation modeled as networks. SNA is now a well-praised methodology in sociology and has been extended to historical research to study relational concepts such as kinship, friendships, and institutions of the past. Social historians leverage their documents to extract relationships between entities—often persons—that they model into networks. Using network concepts and visualization tools, they can make conclusions through structural observations of such networks.

### 2.3.1 Sociometry to SNA

One of Sociology's main goals is to study social relationships between individuals and find recurrent patterns and structures allowing to explain the behaviors of people and groups. Traditional methods try to explain social phenomena using classical social classifications such as age, social status, profession, and gender. However, some criticism emerged that this type of division is often partially biased and comes from predefined categories which are not always grounded in reality [?]. Sociometry is considered one of the bases of SNA and had the goal of redefining social categories through the lens of real social interactions and ties between persons, that sociologists wanted to observe in real conditions. It is in the 1930s that Moreno started to develop this new method by trying to depict real social interactions as a way to understand how groups and organizations were functioning [?]. He developed sociograms as a way to visually show friendships between people with the help of circles representing persons and lines modeling friendships. Figure 2.4 shows one of Moreno's original sociograms to depict friendships in a class of first grades (left). Sociometry tremendously helped disseminate the metaphor of networks to model and understand social structures and phenomena. It was during the 1960s that sociologists and anthropologists took these concepts further and formalized SNA using graphs and mathematical methods, following the emergence of Graph Theory studies in the 1950s by Mathematicians such as Erdős [?]. Sociologists already had structural theories of social phenomena, and they rapidly saw the

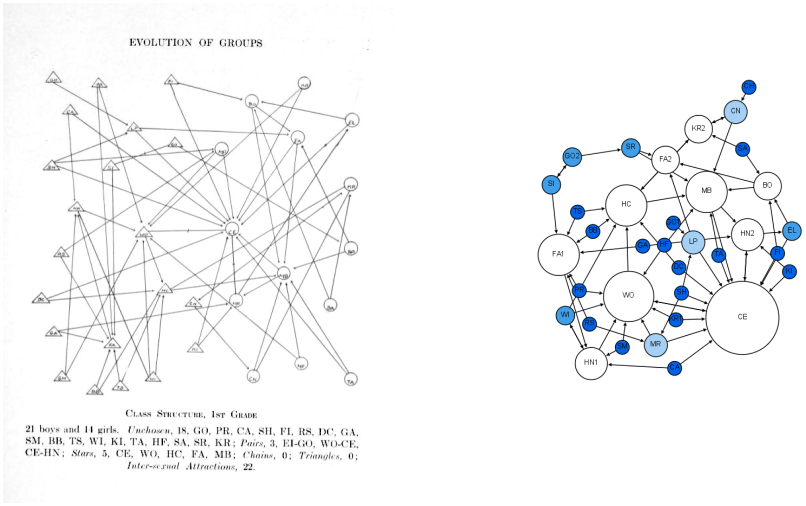


Figure 2.4 – Moreno's original sociogram of a class of first grades from [**?**] (left). The diagram shows 21 boys (triangles) and 14 girls (circles). The same sociogram plot using modern practices generated from Gephi from [**?**]. The color encodes the number of connections incoming.

potential of networks[2] to model social relationships between actors, representing the persons as nodes and relationships as links. Graph theory brought a panoply of concepts and methods to study and describe networks, that sociologists such as Coleman started to codify to use in a sociology setting [?]. Using mathematical and network methods, it was possible to formally describe social relationships to make sociological conclusions grounded in real observations modeled as networks.

### 2.3.2 Methods and Measures

The goal of SNA is to study the structure of a given network to make sociological conclusions. Yet, two distinct methodologies emerged through the history of SNA: the structuralists and the school of Manchester [?, ?, ?].

The structuralists are interested in observing the relational structures and patterns forming a network, to make parallels between them and the social behaviors of actors in real life [?]. They think the positions of the persons in the network and the relational patterns they are part of reflect well the social activities and behavior in real life. Studying those would thus allow them to make interesting sociological conclusions. Accordingly, sociologists in this school usually study organizations and specific groups—such as institutions, companies, families, etc.—and

[2] Graphs and networks refer to the same thing but are often used in different contexts. The term graph is preferred in a mathematical and abstraction setting, while the term network is mostly used when modeling real-world phenomena. We talk about nodes and links for networks and vertices and edges for graphs.

want to explain their functioning through the description of the internal shapes and structures of the networks. Thus, they try to construct networks that exhaustively model all the interactions between the actors constituting the groups, as missing links would misrepresent the reality of interactions.

In contrast, the school of Manchester constituted by anthropologists focuses on studying specific individuals and all their interactions in the different facets of their lives and through time. They typically want to explain certain behaviors and social characteristics of individuals by their relationships and interactions in all their complexity and highlight the influence of different social aspects between them in one's life. One famous example is Mayer's study on austral Africa rural migrants going to cities [?] where he showed that the integration of urban mores and customs were directly correlated to the persons' relationships networks in the city. Xhosa[3] people still interacting with rural people of their village in the city were less changing their customs. This school of thought typically relies on the concept of ego network and more recently dynamic and multiplex networks. Ego networks are networks modeling all the direct relations of one central node—in this case, a person—including the relations existing between the persons of this small network. They typically try to model the different types of relationships of a person, like their family, work, and friendship ties, and study them through time. By studying the ego network structure of someone, sociologists of this school try to leverage explanations on other social aspects of the persons like their social status, job, and gender. It is also common to compare several ego networks to make correlations between the social relationships of individuals and other interesting social categories.

These two methodologies of SNA are often not exclusives and current studies are typically inspired by those two traditions. This is especially true in history where even if historians may want to describe exhaustively a group or institution of the past, they are almost always interested in specific individuals they study in depth.

Furthermore, the two approaches leverage similar network measures and concepts. A myriad of graph measures (e.g., density, centrality, and diameter) and algorithms have been proposed by graph theoreticians and network scientists that social scientists appropriated to describe and characterize social phenomena.

When constructing networks, the first thing sociologists do is to identify the main actors of the network and explain why these actors are the most central, for example by linking it to their profession or social status. Computing the degree—which is the number of connections for a node—distribution is the main straightforward way of doing it, but other more complex measures like centrality have also been developed. Several types of centrality measures (e.g., betweenness, closeness) have been proposed, based on different criteria, as there are several ways

[3]Xhosa people are an ethnic group living in South Africa and talking the Xhosa language. and studied

of defining the more important actors. Some centrality measures highlight actors with the highest number of connections while others highlight people bridging different groups with low interactions. More generally sociologists aim at identifying recurring patterns of sociability between actors. The concepts of dyads and triads counting, which are basic structural patterns of 2 and 3 nodes, give insights into low-level relationships between people. This reflects on Simmel's formal sociology, where he already referred to dyads and triads as a primal form of sociability [?]. More recently, graphlet analysis extended this concept to every pattern of N-entities. Graphlet analysis aims at enumerating every small structure of $N$ nodes composing a network, to understand how people interact at a low level. Graphlets counting shows that graphlets are not found in a uniform distribution in social networks, thus revealing that social networks usually do not have the same structure that random networks. This is a fact well known in SNA. Precisely, entities in real-world networks tend to agglomerate into groups (also called clusters) where entities in the same groups interact more between them than with entities from other groups. From a sociology perspective, it means that people tend to interact and socialize in groups and interact more rarely with other people from outside groups. These groups are often referred to as *communities*, and many algorithms have been proposed to find these automatically.

### 2.3.3 Historical Social Network Analysis

History started to use concepts and methods from SNA in the 1980s [?] in order to criticize quantitative history concepts and results, and to develop historical approaches—like *Microstoria* [?]—that focus on the study of individuals and groups through the lens of their interactions and relationships directly extracted from historical documents. Beforehand, historians were already describing and studying relational structures such as families and organizations with qualitative methods and with classical taxonomies, without studying in depth the relational aspect of these entities. Network research allowed them to model those relational entities more thoroughly using network concepts, thus allowing them to make new observations that it was not possible to see without taking into account the relational aspects of these entities. Observing and describing the structure of the resulting networks allowed historians to make conclusions on sociological aspects of the past, similar to SNA. Since then, HSNA—a term coined by C. Wetherell in 1998 [?]—has been applied by historians to study multiple kinds of relationships, like kinship and political mobilization [?], administrative and economic patronage [?], etc. If these approaches fall under similar critics as quantitative history [?] like leading of trivial conclusions, it still led to classical work and interesting discoveries. One famous example is the study of the rise of the Medici family in Florence in the 15th century by Padgett [?], where he explained the rise of power of this family by their central position in the trading, marriage, and banking networks of the powerful families of Florence. Figure 1.1 shows the different networks of Florence families where we can see the central position of the Medici.

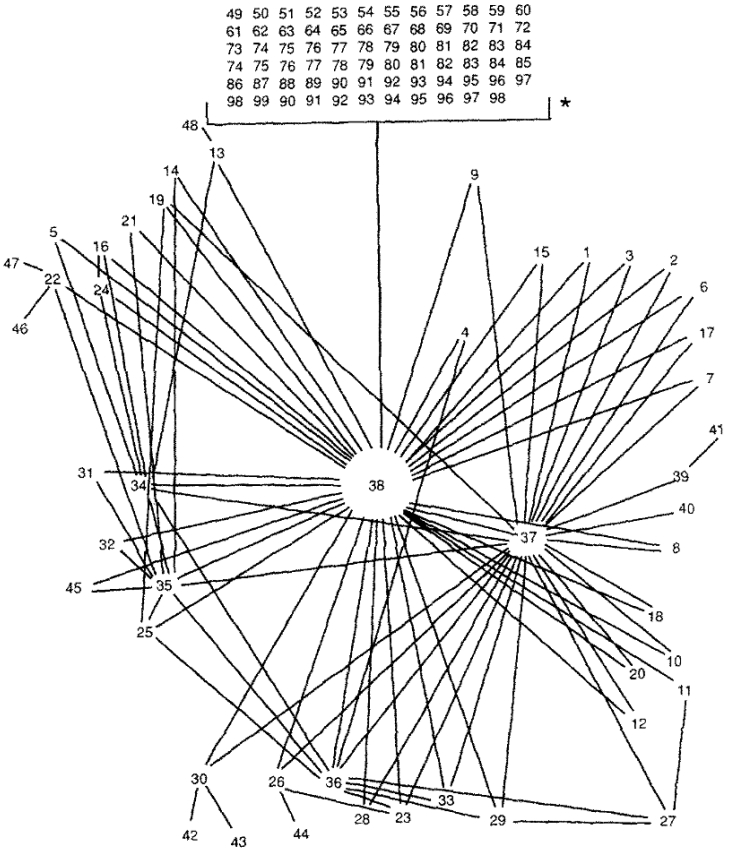


Figure 2.5

Several historians are using and continuously improving the HSNA methods which can be very effective to study relational historical phenomena [**?**]. Moreover, historians rarely rely on a single approach when studying an era or phenomenon, they mix methods and tools from several domains of social and natural sciences with their own practices [**?**, **?**].

### 2.3.4 Network Modeling

Constructing a network from historical documents, which can vary tremendously in their formats and structures, is not a trivial task [**?**]. The most straightforward and well-known approach consists in constructing a social network based on a simple graph $G = (V, E)$ with $V$ a set of vertices representing the actors of interest (very often individuals mentioned in the documents), and $E \subseteq V^2$ a set of edges modeling the social ties between pairs of actors. This allows to have a simple network to visualize and analyze, but it does not always reflect the sociological complexity of information contained in the documents. HSNA network models have evolved over time to better take into account concrete properties of social networks, such as types of actors using labeled networks, the importance

of actors or relations with weighted networks, mixed relationships with multiplex networks, dynamics of relations with dynamic networks. Bipartite networks have been proposed to model relations between two types of entities, such as organization and employees where the relations link employees to organizations but not employees to employees or organizations to organizations. Many social situations or documents can be modeled in these terms (affiliation lists or co-authoring). Multivariate networks, i.e., graphs, where vertices and edges can be assigned multiple "properties" or "attributes", are less used in SNA. These attributes are often considered secondary, the emphasis of SNA being on the topology, its features, measures, and evolution.

Historians, demographers, sociologists, and anthropologists have also been designing specific data models for their social networks, based on genealogy or more generally kinship [?]. For genealogy, the standard GEDCOM [?] format models a genealogical graph as a bipartite graph with two types of vertices: individuals and families. This format also integrates an "event" object but it is diversely adapted in genealogical tools. The Puck software has extended its original genealogical graph with the concept of "relational nodes" to adapt the data model to more family structures and to integrate other social relationships for anthropology and historical studies [?].

## 2.4 Social Network Visualization

Practitioners of SNA and HSNA have always visually depicted their networks for validation and communication purposes, mostly using node-link diagrams. With the increase in average network size and density and the diversity of network models, new visualization techniques have been proposed to represent the diversity of studied networks. Moreover, more and more social scientists are following exploratory approaches using Visual Analytics (VA) tools, to describe more in-depth their data and generate new interesting hypotheses, using interaction and exploration capabilities.

### 2.4.1 Graph Drawing

Sociologists rapidly saw the potential of graphically showing relationships between individuals, to better comprehend the underlying social structure and communicate their findings. Moreno elaborated sociograms to visually show friendships among schoolchildren with circles and lines to respectively show children and friendships ties [?]. This type of representation—commonly called node-link diagram—is the most widely used in social sciences, as it is rapidly understandable and effective for small to medium-sized networks which is usually the norm in social sciences. The most used social network visual analytics software such as Gephi [?] and Pajek [?] are based on this type of representation and allow a fully integrated exploration and analysis with the help of various algorithms. Finding an optimal placement for the nodes is however not that simple as several metrics

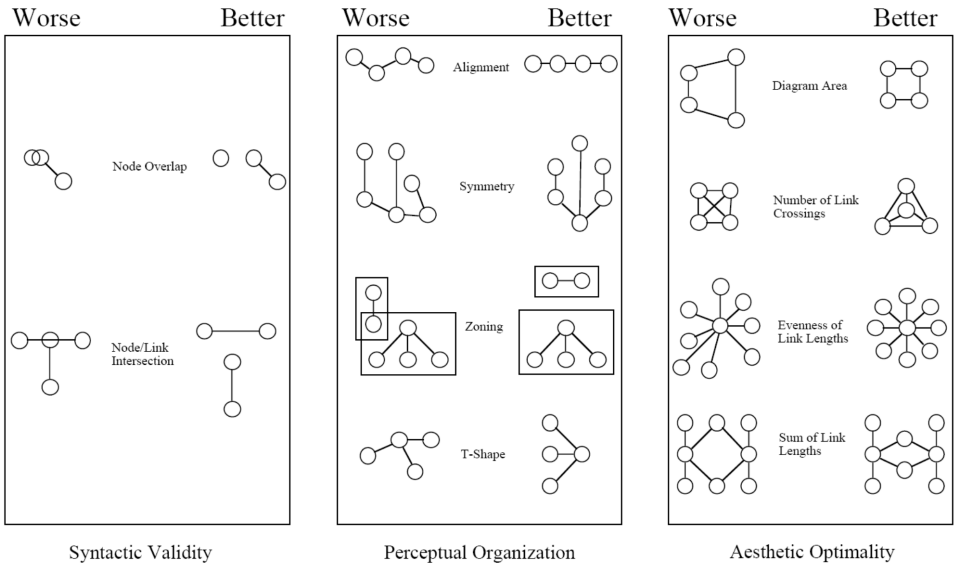


Figure 2.6 – Different criteria are proposed to enhance node-link diagram readability. Image from [**?**]

can be optimized depending on the desired drawing, such as the number of edge crossings, the variance of edge length, orthogonality of edges, etc [**?**, **?**]. Figure 2.6 shows some of these metrics, synthesized by Kosara and al. [**?**]. In Figure 2.4 we can see the difference in readability between the original manual layout (left) and an automatic one (right). Automatic layouts which aim at optimizing readability metrics give clearer diagrams. The number of edge crossings is often considered the most important measure, but finding a drawing with the optimal number of crossings is an NP-Hard problem, meaning that heuristics are needed for most real-world use cases. A large number of algorithms have been designed such as force-directed ones, modeling the nodes as particles that repulse each other and are attracted together when connected with a link that can be seen as strings. Other visual techniques have been proposed to represent networks such as matrices, circular layouts, and arcs, but are less used in social sciences [**?**]. Still, Matrices have been shown to be more effective than node-link diagrams for several tasks such as finding cluster-related patterns, especially for medium to large networks [**?**, **?**].

As social scientists are using more complex network models such as bipartite or temporal networks, more sophisticated representations are needed. The visualization community developed new representations to visualize other network types such as dynamic hypergraphs with PAOHVis [**?**], clustered graphs with NodeTrix [**?**] (illustrated in Figure 2.7), geolocated social networks with the Vistorian [**?**], and multivariate networks with Juniper [**?**]. However, these new network representations take time to be adopted by social scientists who rarely use those.

### 2.4.2 Social Network Visual Analytics

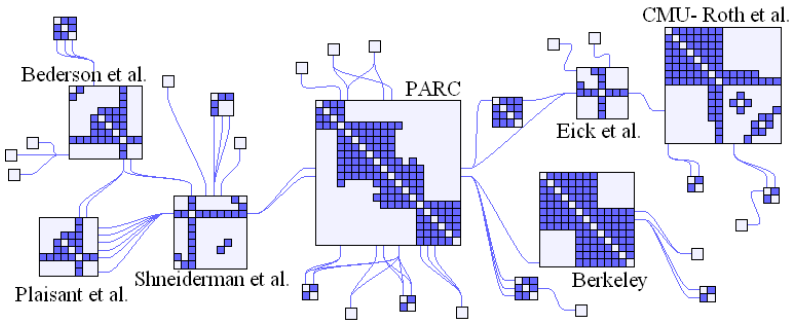


Figure 2.7 – NodeTrix system showing a scientific collaboration social network with clusters. Each cluster is represented as a matrix, Image from [**?**].

Most widely used social network visualization softwares by social scientists are Gephi [**?**], Pajek [**?**], and NodeXl [**?**] which provide node-link diagrams, and allow basic interactions such as selection to explore the network. These softwares usually provides automatic computation of several network measures such as the density and the diameter, allowing users to follow SNA workflows all in the interface. Similarly, Automatic clustering capabilities are provided letting users find interesting community structure in their network. Figure 2.8 presents the Gephi interface showing a clustered social network, where each node is part of a cluster, encoded by color.

More complex VA interface have been proposed to explore social networks with complex interactions and more complex network models such as GUESS [**?**] and the Vistorian [**?**]. These interfaces let social scientists explore their data with other interactions such as filtering, and often propose multiple coordinated views allowing to see the data through different lens. For example, the Vistorian can show the data using a node-link diagram, a matrix, a map, and arc diagrams.

Unfortunately, social scientists are often not trained in computer science and mathematical methods, and most of them have been frustrated by VA tools and by how it was guiding their analysis in predefined ways. Interfaces leveraging automatic data mining algorithms sometimes put users in an awkward position, as they have a hard time interpreting results coming from those black-box algorithms. They usually end up trying several algorithms until they stumble upon a satisfactory enough solution [**?**].

Cleaning and importing the data is also complicated, as the annotation and network modeling process are not straightforward. Thus, social scientists often encounter errors and inconsistencies in the data once they visualize it, that they want to correct. It leads to continuous back and forth between their analysis process inside the VA tool they are using, and their original sources and annotation/modeling process, to correct errors or modify annotations. Interestingly, the network model choice plays a crucial role in the process, as a simple network

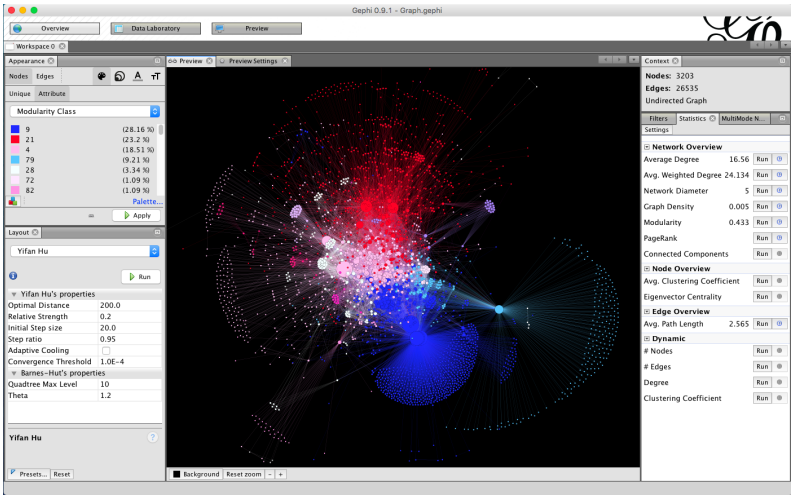


Figure 2.8 – Gephi [**?**] interface. The network is represented with a node-link diagram. Users can interact on the visualization and encode node and links visual attribute (color, size etc.) with network measures computed directly in the interface, such as the node degree, or clustering results.

model representing only the persons (as it is often the case) will make it harder to trace back to the original documents containing the annotations from the network entities. Yet the majority of Social Network VA systems enforce simple network models, making this retroactive process harder. Some interfaces still incorporate data models encapsulating document representations, such as Jigsaw [**?**] which is a VA systems using textual documents as a data model, originally developed for intelligence analysis. It allows an analysis of the documents and their mentions of entities (persons, locations, institutions, etc.) through multiple coordinated views. Using such model allow to rapidly see errors and inconsistencies in the documents annotations, while still following complex analyzes.

Finally, more work is still to be done on social network VA tools, to provide more guidance and power to social scientists while doing their analysis, and to help them to do easier back and forth between their analysis and the annotation, network modeling, and cleaning steps, as they play a big role in the historian workflow.

main