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 Titre, Affiliation	Président ou Présidente
Prénom Nom Titre, Affiliation	Rapporteur & Examinateur / trice
Prénom Nom Titre, Affiliation	Rapporteur & Examinateur / trice
Prénom Nom Titre, Affiliation	Examinateur ou Examinatrice
Prénom Nom Titre, Affiliation	Examinateur ou Examinatrice
Prénom Nom 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 [72, 149]. 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 [43, 127]. 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 [113]—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 [42]. 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"" [19]. Graphically displaying data allows us to leverage our visual human 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

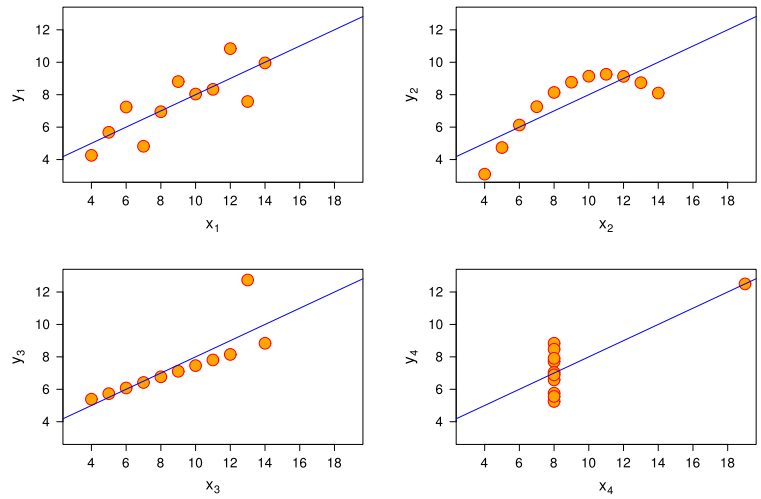


Figure 2.1 – Anscombe quartet. The four datasets have the same descriptive statistics (average, variance, correlation coefficient) but very different structures. Image from [4].

Information Visualization focus on displaying abstract data to amplify cognition and gain insight on real world phenomena [19]. 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 [45], or Snow's dot map of cholera cases in London which showed the proximity between street pumps and cholera infections [135]. 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 [140] and Bertin's publication of Semiology of graphics [8]. 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" [46]. 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 [4] 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* [141] 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" [71]. **??** shows an abstract representation of the VA process.

It is defined around the generation of knowledge using visualizations and models

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.

2.2 . Social History

If Sociology and Anthropology started to use network concepts and methods rapidly in the 1950s, it was not until the 1980s that historians started to use this type of methodology. Yet, historians started to use quantitative methods in the 1960s, with the rise of social history, by extracting information from historical textual documents and studying them with statistical methods. When seeing the potential of SNA concepts for historical purposes, historians started to extract the relational information contained in documents to study historical social phenomena using the power of networks and methods already developed in SNA.

2.2.1 . History, Social History and Methodology

The concept of History is hard to define as its practice and cutomes highly evolved through time [113]. History of a given time can be characterized by the different historical work produces at that time. However, the work of the historians always focused on the collection and study of historical documents, to characterize 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" []. History can be seen through two facets: one is societal, and serves creating a shared story for the country and a sense of unity to its citizens. Antoine Prost says that "it's through history than France thinks itself" (translated from french) [113]. The second facet of history constitute a methodology to describe the past in a rigorous and scientific way, with proofs. For this, historians rely on historical documents that they leverage to infer dated facts about the past (the temporal aspects of conclusions is always central to the historian work). The textual sources are thus at the core of the work of the historians, and having to cite historical documents and previous peers work to new claims is primordial to be considered as rigorous History work. However, even if those two aspects are well characterized (temporal aspect of the work and its relationships to sources), 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 discuss for a long time, until the end of the 1980s. Some historians were interested in historiography [18], but none were going to philosophical and epistemological reflexions of the History discipline. For Lucien fèbvre, philosophising was even constituting a "capital crime" [40, 113].

At the start, history was mainly event-centered, and was focusing in characterizing central figures of the past like rulers and artists or shed light on events which shaped history 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 [14].

In the 1930s, March Bloch and Lucien fèbvre detached from traditional history by creating the "Annales school" (Ecole des Annales) which tried to replace the human as a component of a broader sociological, political and economic system with influences between each other [16]. 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 the real and with a better intelligibility. This school of thought can be seen as one of the biggest milestones for Social History, a branch of 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 [76]. 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

Cliometrics

With the development of statistical methods and more precisely Computer Science, quantitative approaches of History emerged in the 1960s with the aim of analyzing quantitative data directly extracted from historical documents. Using such methods, historians were able to make conclusions based on statistical results on topics such as demography [63] or job distribution. For example, Gribaudi and Blum illustrated a shift in the most widespread professions in France during the 19th century using the data extracted from 50000 marriage acts [55] and using statistical methods.

Unfortunately, quantitative and numeric methods have been criticized by historians for their simplifications and for consuming considerable time while often providing simple results [70, 83]. Trying to understand complex historical phenomena is complicated and modeling the information contained in historical documents into quantitative datasets can rapidly simplify and distort reality. Moreover,

[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.

quantitative historians have been criticized for focusing too much on the data, neglecting the original sources which give the context in which the data has been produced [81]. Guildi and Aritage went as far as criticizing the decrease of interest of historians working in archives [56]. 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). It 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 [69] 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 third wave of computational social sciences [81]. The term have gained popularity since the 2010s and refer to "research and teaching taking place at the intersection of digital technologies and humanities. DH 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). DH also studies the impact of these techniques on cultural heritage, memory institutions, libraries, archives and digital culture." [?]. If the first waves of computational social sciences focused a lot on mathematical computations such as hypothesis testing and correlation computations to make conclusions, DH focuses also on the teaching, communication and exploration of humanities datasets and concepts, through the help of 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."" [?] Digital History therefore focus a lot on 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

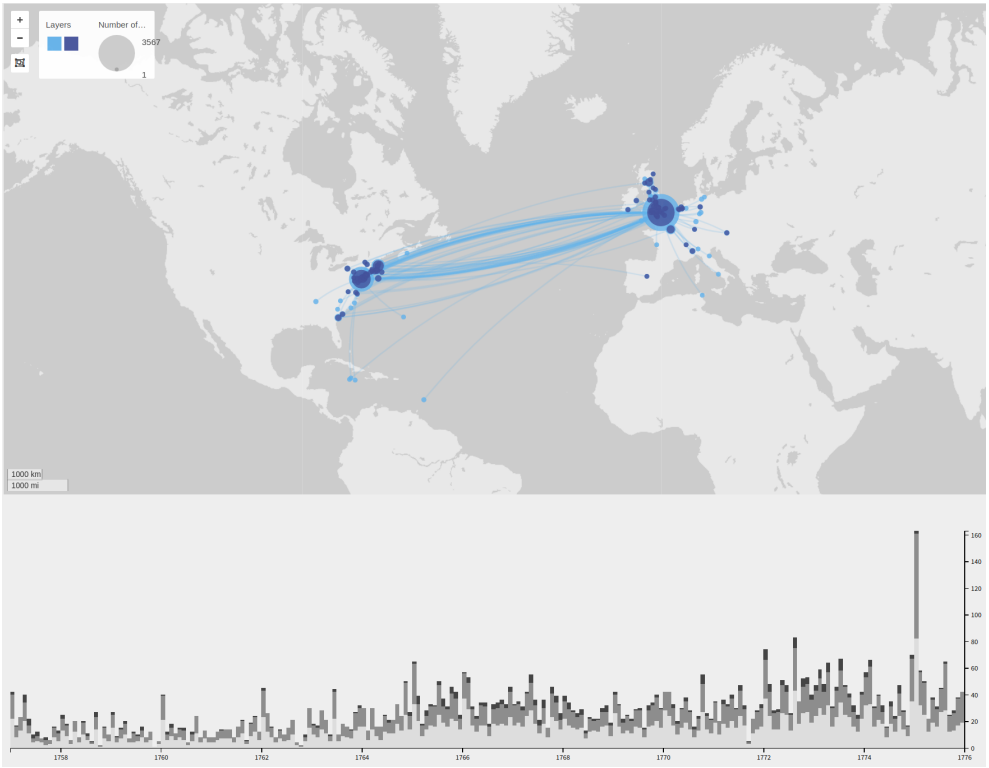


Figure 2.2 – Correspondence letters of Benjamin Franlin and his close relationships, using a map and an histogram, accessible online on the republic of letter website.

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 2.2), 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 as with the Venice Time Machine project [69] 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. A lot of projects which claim themselves as Digital History also leverage network methods and concepts [**?**], such as 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 network analysis for historical research in **??**.

2.3 . Historical Social Network Analysis

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 concept of SNA emerged in sociology in response to traditional 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. For example, the socio-economic position of people living in a small city could be explained by their age, demographics, and family status which are traditional

social categories. 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 [95]. He developed sociograms as a way to visually show friendships between people with the help of circles representing persons and lines modeling friendships. Figure 2.3 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 [37]. Sociologists already had structural theories of social phenomena, and they rapidly saw the 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 [21]. 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 [39, 43, 85].

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 [79]. 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 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.

[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.

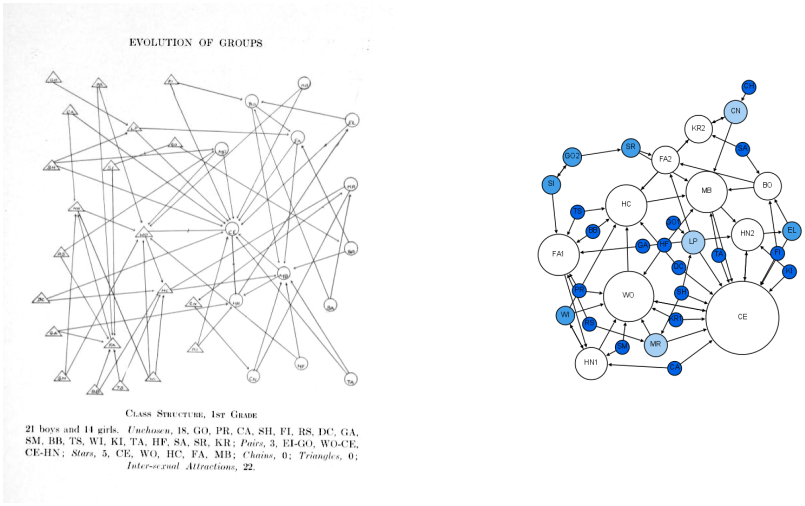


Figure 2.3 – Moreno's original sociogram of a class of first grades from [94] (left). The diagram shows 21 boys (triangles) and 14 girls (circles). The same sociogram plot using modern practices generated from Gephi from [53]. The color encodes the number of connections incoming.

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 [86] 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

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

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 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 [132]. 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 [149] in order to criticize quantitative history concepts and results, and to develop historical approaches—like *Microstoria* [49]—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 has been applied by historians to study multiple kinds of relationships, like kinship and political mobilization [84], administrative and economic patronage [97], etc. If these approaches fall under similar critics as quantitative history [80] 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 [105], 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 2.4 shows the different networks of Florence families where we can see the central position of the Medici.

Several historians are using and continuously improving the HSNA methods which can be very effective to study relational historical phenomena [72]. 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 [105, 109].

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 [2]. 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,

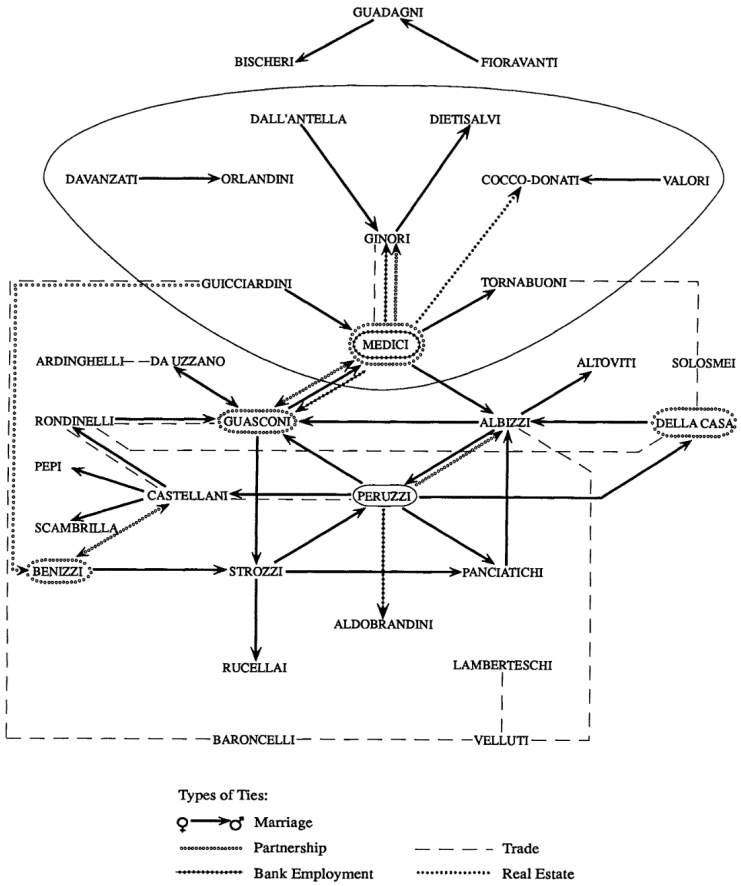


Figure 2.4 – Marriage, partnership. trading, banking, and real estate networks of the powerful families of Florence from [105]. We can see the central position in the network of the Medici Family.

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 [59]. For genealogy, the standard GEDCOM [47] 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 [58].

2.4 . Social Network Visualization

Practitioners of SNA and HSNA have always depicted visually 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 [94]. 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 [5] and Pajek [98] 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 can be optimized depending on the desired drawing, such as the number of edge crossings, the variance of edge length, orthogonality of edges, etc [24, 75]. Figure 2.5 shows some of these metrics, synthesized by Kosara and al. [75]. In Figure 2.3 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 [88]. 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 [**?**, 48].

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 [142], clustered graphs with NodeTrix [64] (illustrated in Figure 2.6), geolocated social networks with the Vistorian [128], and multivariate networks with Juniper [102]. However, these new network representations take time to be adopted by social scientists who rarely use

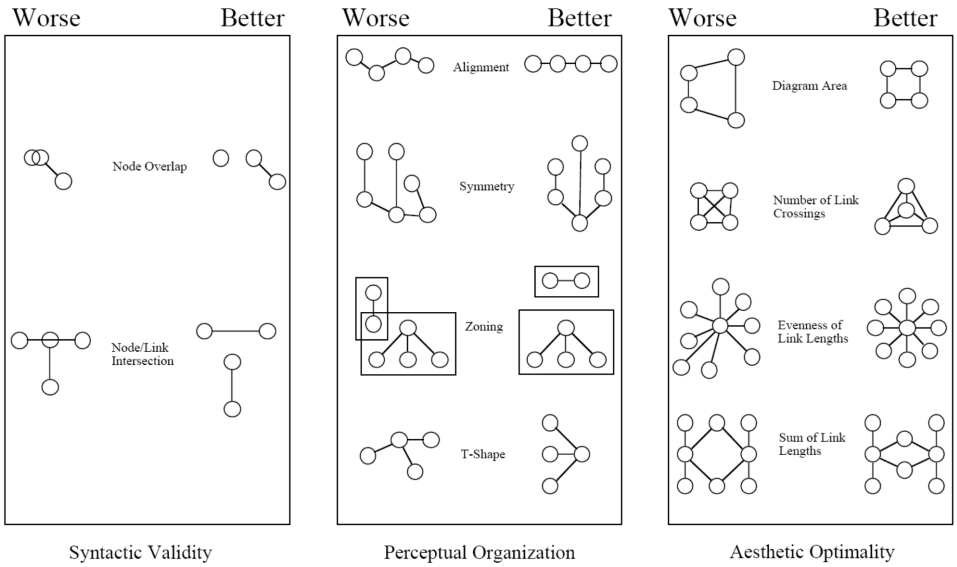


Figure 2.5 – Different criteria are proposed to enhance node-link diagram readability. Image from [75]

those.

2.4.2 . Social Network Visual Analytics

Most widely used social network visualization softwares by social scientists are Gephi [5], Pajek [98], and NodeXl [133] 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 2.7 presents the Gephi

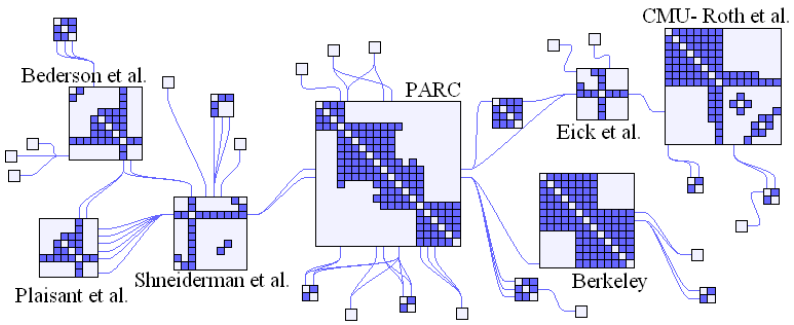


Figure 2.6 – NodeTrix system showing a scientific collaboration social network with clusters. Each cluster is represented as a matrix, Image from [64].

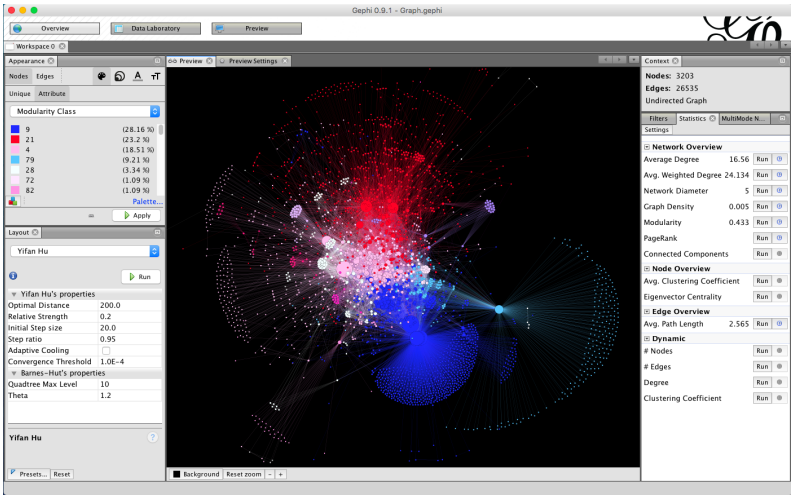


Figure 2.7 – Gephi [5] interface. The network is represented with a node-link diagram. Users can interact on the visualization and encode node and links visual attribute (color, size etc.) with network measures computed directly in the interface, such as the node degree, or clustering results.

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 would like to correct. Historians thus always have to go 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 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 [136] 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.

3 - HSNA Process and Network Modeling

We describe in this chapter the HSNA workflow followed by social historians, to shed light on their process and summarize recurring pitfalls to identify how VA could help them in this process. Specifically, we discuss in depth the network modeling step, as the choice of the network model influence the overall process, especially the possibilities of the analysis. Most HSNA practitioners report on their findings concerning the network they constructed from their sources, but few highlight their process which led to these conclusions from the raw historical documents. Similarly, VA tools always focus on the analysis part, once the network have been constructed, without helping historians in the previous steps. However, the data collection, encoding, and transformation steps are crucial and can introduce lots of bias and distortion on the final data if not done correctly. This is especially true for social history where historical documents can lack structure and can be hard to parse, and where historical claims should be traceable to the original sources. We therefore describe the HSNA workflow split into 5 steps and characterize recurring pitfalls which can occur in each step. We also discuss in depth the network modeling step, as social historians can model their documents with various models which have an impact on the representation of the social relationships, traceability to the documents, and simplicity of usage.

This chapter is an updated version of an article presented at the VIS4DH workshop of the IEEE VIS: Visualization Visual Analytics Conference 2022, and published in IEEE Explore [112]. It was a collaboration with Nicole Dufournaud, Pascal Cristofoli, and my supervisors Christophe Prieur and Jean-Daniel Fekete. I participated in the discussions, elaboration of concepts, and writing of the paper.

3.1 . Context

Tools for social network visualization tend to ignore the context in which the networks are produced, where they come from, and the workflow that led from their origin (e.g., documents, polls, interviews, web scraping) to their network form. Yet, practitioners of social history need to generate many networks from the same documents/sources to visualize and analyze them. In this chapter, after describing and characterizing the workflow of Historical Social Network Analysis [149] from our collaborations with social historians, we explain why and how effective tools for supporting this process should model social networks in multiple steps to support three essential principles: *traceability*, connection to *reality*, and *simplicity*. These principles emerged from joint experiences as historians and computer scientists while collaborating on multiple projects.

Social historians' goal is to characterize socio-economic phenomena and their

dynamics in a restricted period and place of interest and to see how individual people of that time lived through those changes. For this, they rely on historical documents such as conversational letters, censuses, and marriage acts. They usually extract qualitative and quantitative information from an identified corpus of documents, to then make conclusions on interesting socio-economic topics such as migrations, business dynamics, education, and kinship. For doing this, historians can apply HSNA methods, by modeling the social relationships between a set of entities—usually individuals—into a network. Historians therefore collect documents, annotate them, construct a network from the annotations that they finally analyze and visualize to validate or find new hypotheses. Unfortunately, the process is often linear, and it is common that, when visualizing their network, historians spot errors and inconsistencies in the annotations that they could have fixed if the process was iterative.

Moreover, historical documents are often complex and the annotation and modeling process can be done in many ways. Several network models have been proposed ranging from simple and specific ones like co-occurrence networks to more general and complex ones such as multilayer networks and knowledge graphs. Simple models allow answering specific questions and are easy to manipulate but are often too simplistic and may distort the information contained in the documents. Moreover, they often break the traceability from the analysis to the original documents, making the communication of findings less reproducible and the process of cleaning the annotations complicated. Indeed, errors and mismatches often occur in the annotation process, for example, due to entity disambiguation problems. On the contrary, too complex models are complicated to visualize and analyze, and historians do not always have the tools to create them properly. In this chapter, we propose to model historical datasets as bipartite multivariate dynamic networks, where both persons and documents are modeled as nodes with attributes. While this model is simple enough for creation and inspection, it allows tracing back the entities of the network to the original sources for a continuous annotation process and still accurately models the social relationships mentioned in the documents. Historians can therefore use this model to simultaneously find errors and inconsistencies in their annotation process—allowing them easier back and forth between the annotation and analysis steps—while starting a first analysis and exploration of the data to answer their sociological questions. The traceability to the original sources also makes the communications of findings more replicable and transparent.

3.2 . Related Work

Since we already elaborated on the related work of SNA, HNR, network modeling, and social network visualization in chapter 2, we only discuss in this section the related work concerning historians' workflow and methodology descriptions.

The essence of the historical discipline is based on a critical approach of sources

and involves considering peers' work. Traditional approaches to history often focus on the construction of a narrative, without necessarily adopting a systematic and problematized approach to the exploitation of original sources. Social history and the "Annales School" proposed a new approach to history, by trying to describe and characterize socio-economic phenomena of the past by rigorously extracting information from historical documents and making conclusions from them.

With similar aims, Glaser and Strauss developed the "Grounded Theory" [50] as a methodology for the humanities to build hypotheses and theories by solely studying and categorizing real-world observations, without starting from prior knowledge and predefined categories. Later on in the 1960s, quantitative methods started to be used in history, providing statistical and later computer-supported tools to aid historians in grounding their analysis in mathematical models and results. Unfortunately, the lack of methodology and understanding between the two worlds led to many criticisms by historians pointing to using wrong metrics, simplifying categories, and disconnections between the original documents and analysis [70, 82]. Quantitative history has been showed to be useful when used properly and when not focusing only on numbers, and several books have been published on how to efficiently use statistical methods such as summarizations, correlations, statistical distributions, statistical testing, time series etc. [66, 81]. Similarly, the use of network science for historical aims increased in recent years, and a lot or resources exist on how to use network methods and measures for historical research [72, 80].

However, little work has been done on describing and formalizing the process before the analysis part for a quantitative and network research workflow. Indeed, if it is central to know how to manipulate statistical and network concepts and methods when following this kind of methodology, it is as important if not even more to follow a correct and rigorous workflow to generate the data we plan to analyze beforehand. The process to generate a clean quantitative or network dataset from historical sources is difficult and requires several data acquisition, annotation, and cleaning steps. Social analysts are not always trained on how to do these steps effectivity, which can lead to errors, inconsistencies, and mismatches between the chosen data models and the historical questions [2]. Karila-Cohen and al. provide some advice on how to annotate historical documents with the aim of using quantitative methods [70] and prone that the annotation and analytical processes should not be dispatched between several persons, as both usually influence each other. Dufournaud describes her workflow in depth when studying the socio-economic status of women in France in the 16th and 17th centuries, which she splits into three steps: *data collection*, *data processing*, and *data analysis* [34]. She provides the tools and methodology she used to annotate her data, providing transparency on her historical analysis and methodological resources. Cristofoli discusses the network modeling problem when following an HSNA and highlights the fact that the same historical documents can be modeled in different ways [23]. Historians should be aware of this and choose a network model which fits their analytical

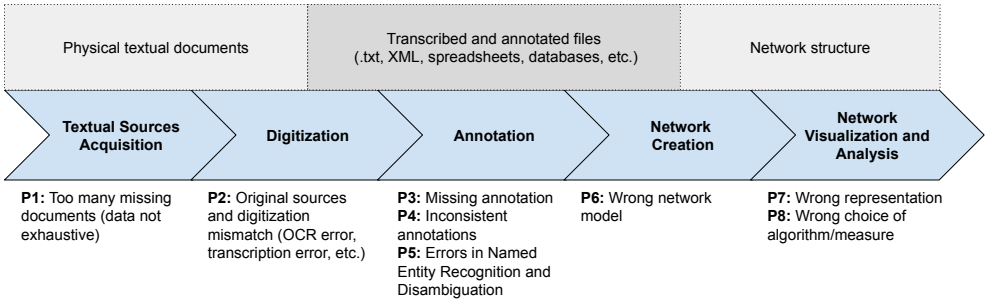


Figure 3.1 – HSNA workflow is split into five steps: textual sources acquisition, digitization, annotation, network creation, network visualization and analysis. We list potential pitfalls for each step.

goals.

3.3 . Historical Social Network Analysis Workflow

From the literature and our own projects of HSNA we conducted during the last years in collaborations with historians, we propose an HSNA workflow divided into 5 steps: *textual sources acquisition*, *digitization*, *annotation*, *network creation*, and finally *visualization and analysis*. The workflow is presented in Figure 3.1 along with potential and recurrent pitfalls.

3.3.1 . Textual Sources Acquisition

Historians' first step is gathering a set of textual historical documents mentioning people with whom they will have social ties. For this, they usually take documents from a specific source—such as a folder from a national or local archive—and restrict them to a period and place that they want to study. They also often restrict themselves to one document type—such as marriage or notary acts—to focus the analysis on one or few types of social relationships that they want to understand in depth. However, one rule of the historian's method is to crosscheck from multiple sources, so an initial corpus is often extended with another set of related sources. Once they restricted their search to a set of documents, a time, and a geographic area, they try to exhaustively find all the documents matching the desired properties, as **missing documents can result in uncertainty in the network structure and therefore the sociological conclusions (P1)**.

3.3.2 . Digitization

Digitization consists in converting the sources into a digital format. This step can be skipped for the most recent periods where many documents have been produced digitally or can be scanned and well digitized through optical character recognition (OCR), allowing to tremendously ease the storage, indexation, and annotation of the documents. However, before mid 20th century, most historical

primary sources are stored in archives in paper format and need human work to be digitized. **Mismatches between the original documents and the transcription can occur for old and recent documents (P2)**. However, if OCR tools are more and more efficient in English and highly used languages, historians can work with old documents written in old or extinguished languages and with atypical writings (e.g., Fraktur handwriting and typefaces for German in the early 20th century). Therefore, OCR tools are often unusable in social history and digitization remains an expensive and sometimes highly skilled process.

3.3.3 . Annotation

Annotation is the process of finding and extracting useful information from the documents concerning the persons, their social ties, and any useful information for the historian. This extra information can concern the persons (their age, profession, sex, ethnicity, etc.) and their social relationships (type, date, place). It encompasses named-entity recognition (NER) as well as their resolution. Historians also sometimes annotate information on other entities mentioned in the documents, such as art objects or administrative entities. Usually, historians have a first idea of what they want to annotate in the data as they already explored the documents beforehand and have knowledge of their subject of study, with hypotheses they want to explore. It is however common they can change their mind through the annotation process, by reflecting on what they found in the documents. Unfortunately, this can produce **missing annotations (P3)** and **inconsistent annotations (P4)** at the end of the process if annotators are not careful. This task can also be challenging and the choice of annotations has an impact on the final network. Historians also face ambiguity in the process, as several persons and entities (like cities) can have the same name (homonyms), refer to a place name that has disappeared (street name or city), or to an ambiguous person (e.g., John Doe). They, therefore, have to follow a NER and resolution/disambiguation process to identify entities in the sources and disambiguate them across several documents. Entity resolution has always been a problem in social history—as it is more generally in text analysis, where typical groundwork consists in crossing information about the same entities from different heterogeneous sources. However, errors in the disambiguation process can lead to important distortions in the final network structure and properties [31], e.g, people connected to the wrong "John Doe".

Historians usually carry out this process manually but can also use automated methods and refine the results themselves later. Unfortunately, **errors are common in this step as automated methods do not provide perfect accuracy, nor doing it manually given the lack of global information (P5)**.

The Text Encoding Initiative (TEI) [22] is an XML vocabulary and a set of guidelines typically used to encode and annotate documents, and the events happening in these documents (unclear parts, gaps, mistakes, etc.). It is also used for historical texts and to generate social networks [35, 128]. Unfortunately, the guidelines are not meant to define a canonical annotation and different persons can

interpret the guidelines in different ways, leading again to inconsistent annotations of corpora (P4) and to errors or distortions in social networks derived from these annotations.

3.3.4 . Network Creation

Historians construct a network from the annotations of the documents. Usually, all persons mentioned are annotated and will be transformed into network nodes (vertices). Additional information such as their age, profession, and gender can be stored as node attributes. How the network's links are created is not as trivial and can vary from project to project [2]. The most straightforward approach is to create a link between every pair of persons mentioned in one document, thus forming a clique motif. This is a simplistic heuristic as social relationships can be quite complex, involving more than two persons who can have different roles in the relationship. The choice of the network model has a major impact on the future analysis and **may add bias if chosen loosely (P6).** More complex models have been proposed in the literature such as weighted, dynamic, bipartite, and layered networks.

3.3.5 . Network Analysis and Visualization

Once historians have constructed a satisfactory network, they start exploring and analyzing it with visualization and quantitative methods. The final goal of HSNA is to find interesting patterns and link them to social concepts to gain high-level socio-historical insights [43, 149]. Usually, historians start to visualize their network to visually confirm information they know, then to potentially gain new insight with exploration. Representations need to be chosen wisely given the network as lots of techniques and tools exist for social network visualization. **Some insight may be seen only with some specific visualization technique (P7)**. To test or create a new hypothesis, historians typically rely on algorithms and network measures. Lots of network measures have been developed like modularity, centrality, and clustering coefficient that social scientists can leverage to make conclusions [127]. Similarly, social scientists can use data mining algorithms to highlight interesting and potentially hidden structures in the network, e.g. by using clustering algorithms revealing group structures [15]. **However, they have to interpret the results carefully (P8)** as some algorithms act as black boxes and some measures are hard to interpret, with unclear sociological meaning (e.g., centrality). Typically, particular patterns and measures values in the network could have different potential sociological meanings. If we take as an example betweenness centrality which measures the number of times a node appears in the shortest path of every pair of existing nodes, individuals with high values usually highlight positions of power as they communicate with different groups. However, it can also be interpreted as a position of vulnerability in other contexts such as during periods of wars and repressions, as in the study of Polish social movements in the 20th century by Osa [104] where she shows persons with high betweenness centrality

values are more targeted for repression in certain periods. Social scientists, therefore, have to be careful when interpreting network measures and take into account the globality of their sources when interpreting the network they constructed.

3.4 . Network modeling and analysis

Historians typically construct one or several networks from their annotated documents that they will visualize and analyze to validate or find new hypotheses. As the processing steps of the workflow are often not transparent (digitization, annotation, network modeling), it can be difficult for the reader of an HSNA study to understand how the network has been constructed, what it represents, and to trace back the network entities to the original sources [34]. Moreover, visualizing the network very often highlights errors and artifacts of the annotations, along with potential mismatches between the network model and the analysis goals. Historians then have to correct or change their annotations, even though it is a very tedious and demanding process to repeatedly switch back and forth between the network and the annotated documents. Several network models make the task harder as they do not directly represent the documents, and it is thus difficult to relate a network entity to a specific document and annotation. Therefore, we believe that more visual analytics tools should support social scientists in annotating and modeling their documents to make the HSNA process less linear by allowing easier back and forth between the annotation, modeling, and visualization steps. Network models satisfying *traceability*, *reality* and *simplicity* properties would mitigate those problems by allowing to navigate more easily between the network and the documents while still modeling well the social relationships mentioned in the sources and being easy enough to visualize and manipulate for analytical and cleaning goals.

3.4.1 . Network Models

Currently, historians use various network models depending on their knowledge of network science, the content of their documents, the schema of their annotations, and the analysis they plan to make. We describe here the most used network models in HSNA along with more recent ones:

- **Simple Networks [149]:** According to their research hypotheses, historians select and merge document information to build a specific relationship between individuals. They analyze this simple network structure with SNA tools and produce network indicators and node-link visualizations. It is often difficult to connect the results to the original sources.
- **Co-occurrence networks [123]:** Only the persons are represented as nodes, and two persons are connected with a link when they are mentioned in the same document (or section). This is a simple model and one of the first to have been used in SNA and HSNA. The major drawback of this model is that it does not take into account the diversity of social relationships, as every link is

identical. It can work well when only one type of social relationship is studied like a friendship network [95]. However, historical documents rarely mention only one type of relationship and this model is thereby very limiting for HSNA.

- **Multiplex Unipartite Networks [38]:** Only the persons are represented as nodes, and links model social ties between two persons. Links can have different types representing different types of social relationships. It allows modeling more complex social relations where people can have various social ties e.g. as parents, friends, and business relationships. However very often several possible representations for the same data exist as projections are often applied to the original documents to get this type of model. One of the main drawbacks of this model is that it creates parallel edges that are hard to visualize.
- **Bipartite (also called 2-mode) Networks [58] :** Nodes can have two types: persons and documents in this network model. A link refers to a mention of a person in a document and can thus only occur between persons and documents nodes. Usually, links are not typed and only encode mentions. More recent analyses in HSNA encode the *roles* of the persons in the documents as link types [25]. This network model is more aligned with the original sources and allows following an analysis through the original documents themselves and not through concepts. For example, the GEDCOM format introduces the concept of "family" that ties together a husband, spouse, and children with different link types. However, the concept of family can have different meanings across time and cultures, meaning that GEDCOM adds a conceptual layer instead of grounding the network to concrete traceable documents and events (e.g., no marriage but birth certificates).
- **Multilayer Networks [87]:** in these networks, each node (vertex) is associated with a *layer* l and becomes a pair (v, l), allowing to connect vertices inside a layer or between layers. These advanced networks have received attention from sociologists [26] and historians [144], but they are complex. The meaning of a layer varies from one application to another; it can be time (years), type of documents, the origin of sources, etc. They, therefore, offer many (too many) options for modeling a corpus, and visualizing it, with no generic system to support historians for taming their high complexity.
- **Knowledge Graphs (KG) [65]:** they represent knowledge as triples (S, P, O) where S is a *subject*, P is a *predicate*, and O is an *object*. Everything is encoded with these triples using controlled vocabularies of predicates and rules known as *ontologies*. KG is popular for encoding knowledge on the web, including historical knowledge. However, it is notoriously complex to encode documents using KG due to the complexity of the format and the wide choice of possible ontologies. Most historians are unable to understand KG and even less to use it for annotating a corpus. Since KG are generic, they need complex transformations to be visualized, with no generic system to support historians in taming their high complexity.

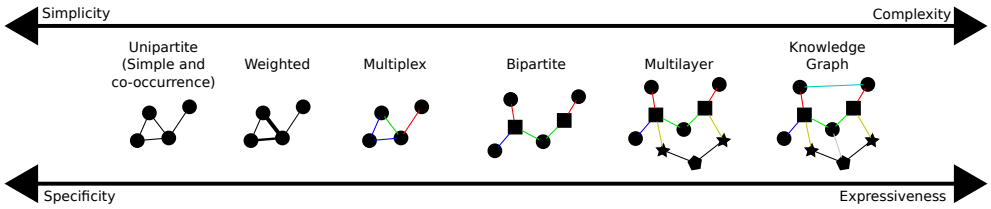


Figure 3.2 – Schematic representations of Different network models used for analyzing historical documents, ordered by complexity and expressiveness

We argue that historians should aim to model their networks simply enough to be manipulated by them, in a way that entities can be traced back to the sources, and expressive enough to model accurately the social reality of the documents—i.e., having those three properties: *simplicity*, *traceability*, and *reality*.

Currently, most digital historical projects use unipartite networks (simple, co-occurrence, and multiplex) that are simple and allow answering specific questions, but they do not capture all the complexity of the documents, and social scientists may miss important patterns. For example, modeling only co-occurrences of persons in documents remove the variety of social relationships these mentions can refer to. Moreover, since documents are not explicit in the unipartite model, it is hard to trace the network entities back to the sources: the traceability property is not satisfied. On the other side, multilayer networks and KG allow to model documents as entities and express complex relationships between various other entities they mention. These models can be very expressive but are challenging to use for historians, especially without guidelines; without *simplicity*, the *traceability* and *reality* properties can be hard to achieve. Moreover, they are difficult to visualize and analyze, especially for social scientists.

Figure 3.2 shows a schematic representation of the different network models, ranked on simplicity/complexity and specificity/expressiveness axis.

3.4.2 . Bipartite Multivariate Dynamic Social Network

Historical documents are well modeled by bipartite multivariate dynamic networks with roles, which have the following properties:

Bipartite: There are **two types of nodes**, persons and documents (or events). An event, such as a marriage, is most of the time witnessed by a document, and we refer to them interchangeably as events and documents. Events considered in the network can be of the same sub-type, such as contracts, or of multiple subtypes, e.g. for genealogy: *birth certificates*, *death certificates*.

Links and Roles: A link models the mention of a person in a document. **Each link has a type corresponding to the role of the person in the document**. For a marriage act, the roles include *wife*, *husband*, *witness*. This is a key aspect of our model since it clarifies the relationship between the persons within an event.

In contrast, Jigsaw [136] does not consider the roles.

Multivariate: Each entity of the model can have attributes, that give additional information. Person nodes are referenced by a key that reflects the disambiguation process. They can have general information (standardized name, gender, birth date). Documents are also identified by a key, e.g., an archive reference. The associated event can have a date, sometimes a location, and potentially other information. Links can also carry information to describe contextual properties (activity, residence, etc.).

Geolocated: Events should have a location when it makes sense, ideally with the longitude and latitude.

Dynamic: Events are always dated. We rely on this date since it encodes the social dynamics of the network.

One of the main benefits of this model is that the document nodes represent both the physical documents and the events the documents refer to. For example, concerning marriage acts, the document nodes represent both the physical documents with their texts and also the marriage events with their characteristics modeled as attributes (time, location, etc.). Therefore, social historians can use this model to store, process, and clean their original documents and follow an analytical workflow with the same representation. This model is *simple* enough to manipulate and visualize for historians and allows tracing back every entity of the network to the documents according to the *traceability* principle. Still, the network preserves the *reality* of the social relationships mentioned in the sources as no projection or transformation is applied.

3.4.3 . Examples

We discussed with four experienced historians collaborators at different steps of their HNSA workflow about their annotation process and how they wanted to model their data into a network. They all work on semi-structured historic documents, mentioning complex relationships. We provide more details in the following:

1. Analysis of the social dynamics from **construction contracts in Italy in the 18th century [25, 101].** The corpus is made of contracts for different types of constructions in the Piedmont area in Italy. People are mentioned under three different roles: *Associates* who are in charge of the construction, *Guarantors* who bring financial guarantees, and *Approvers*, who vouch for the guarantors. Documents contain information about the building site, the type and materials of constructions, and the origin of the people.
2. Analysis of migrations from the **genealogy of a french family between the 17th–20th centuries** [unpublished work]. The corpus is made of family trees referring to several document/event types: birth and death certificates, marriage acts, military records, and census reports. The roles are different for each event type and consist of *children, father, mother* for the birth events, *deceased* for the death event, *spouse* and *witnesses* for the marriages, and *family member*s for the census events.

3. Analysis of migrations from Spain to Argentina through the **marriage acts at Buenos Aires in the 17–19th centuries [96, 122].** The corpus is made of summaries of marriage records that mention the spouses and the witnesses of the wedding. The origin, date of birth, and parents' names are specified for both spouses.
4. Socio-political analysis of **migration of ethnic Germans from communist Romania to West Germany in the 20th century (ongoing work) [32].** The corpus is made of administrative forms that mention persons requesting to migrate, along with the persons they want to join, and the administrative persons of the ministry in charge of the forms. The family members of the aspiring migrants are also mentioned in the forms, with their respective dates of birth.

We compare what would be the resulting networks for the three first examples (the example #4 is still in the phase of data acquisition) when modeling the data with the three most frequently used network models in HSNA: co-occurrence, multiplex unipartite, and bipartite networks. We also encode important information from the document as network attributes. We do this for one given document for each dataset. The results are shown in Table 3.1.

As shown by Cristofoli [23], we can clearly see the co-occurrence model removes the complexity of the social relationships and only shows an abstract "proximity" between individuals. Unipartite projections allow producing meaningful networks which model well the diversity of relations that can link several people. It especially models well simple relationships such as parenting ones as in example #2. However, it produces distortions for more complex relationships involving more than two persons, as in example #1 where people can either be mentioned as associates, guarantors, and approbators in the documents. Associates should probably be linked together with *associate* links, but the *guarantors* and *approbators* relationships are more complex to model. Approbators could be linked to the associates, the guarantors, or both. The three ways of modeling this type of relationship make sense but can lead to very different network shapes and analysis results. Historians thus have to decide on a transformation among several possibilities, which will probably distort the social reality of the relationships.

Moreover, projections add ambiguity in retrospect of the original documents, as it becomes impossible to trace back one link to one specific document, as the same link could potentially refer to several ones [23].

Finally, these examples show that when working with multivariate networks, using projections to create unipartite networks brings a duplication of information. Indeed, if a document mentions information like a date that we model as an attribute, we can store it as a document node attribute using a bipartite model. However, when projecting the network this information appears in the links as many times as there are persons mentioned in the document minus one and often more. For example, in the example #1 in Table 3.1 the time is stored in $\sum_{i=1}^{4} i = 10$

links in the co-occurrence model and in 9 links in the multiplex unipartite model while it is only stored once as a document node attribute in the bipartite model.

3.5 . Applications

Several tools have been designed for visualizing dynamic bipartite networks that can also be considered dynamic hypergraphs [108, 142], but few incorporate attributes. Moreover, the vast majority of visual analytics tools are solely focused on the analytical part of the data, meaning that the link between the original documents and the hypergraph abstraction is often broken. Social scientists therefore always have to do many back and forth between the visual analytics tools and their original documents and the annotation/modeling processes. More visual analytical tools should thus incorporate the textual documents in their data model similarly to Jigsaw [136], as it would allow tracing the entities of the network back to the original documents more easily. Mechanisms to clean/modify the annotations and reflects on the network modeling process directly in the analytical environment could also ease the social scientists' workflow loop. It would allow them to directly clean errors and inconsistencies in the annotations and propagate them in the visual analysis workflow. For example, the Vistorian [128] now lets users modify and clean their data in a table format if they see errors or inconsistencies.

3.6 . Discussion

Most tools for social network visualization focus solely on the visualization and analysis steps, without considering the whole historical data analysis process, preventing researchers from going back to the original source, and supporting the social analyst in the annotation and modeling steps. We think visual analytics tools helping social scientists annotate and model their data with *reality*, *traceability*, and *simplicity* principles in mind are essential to conducting socio-historical inquiries with limited friction, realistic training, and scientific transparency. Concerning the network modeling step, bipartite multivariate dynamic networks model well the majority of structured historical documents such as marriage acts, birth certificates, and business contracts as these documents refer to specific events (birth, marriage, transaction, etc). The document nodes, therefore, represent both the textual documents and the specific events. This dual representation works well for semi-structured documents but could be more limiting for other more literary documents. Moreover, structured documents can also provide information about other relationships not directly linked to the main event. For example, marriage acts sometimes refer to the place and date of birth of the spouses with the names of the parents. This information relates to the birth of the spouses and not the marriage specifically. In that case, social historians can either ignore this type of information in the annotation process or encode it with specific roles (*husband's*

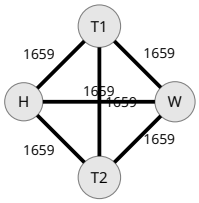
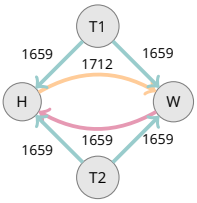
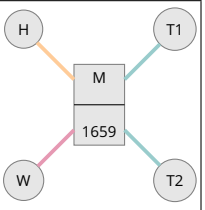
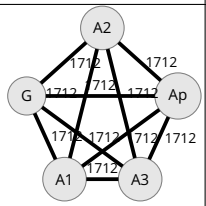
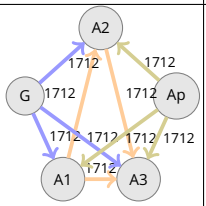
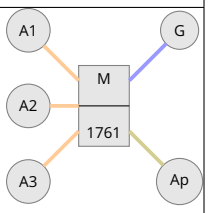
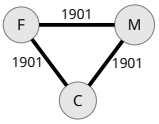
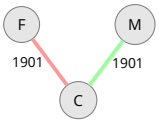
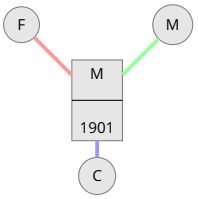
Original Document	Co-occurrence	Unipartite representation	Bipartite
20-4-1659 : Capitán Alonso MUÑOZ de GADEA , con Da. Francisca CABRAL LEAL de AYALA . Ts.: Agustín Gayoso , y Juan Guerrero. Al margen: "fue Oficial Real" , (f. 9v). Husband Wife Witness			
1712: Construction of a church in Torino. Associates: Bellotto G, Bello P.M, Bello G. Guarantor: Astrano G.A. Approbator: Corte A. Associate Guarantor Approbator			
Du dix-neuf fevrier mil huit cent quatre-vingt quatre, à six heures du soir. Acte de naissance de Dufournaud Alexis, enfant de sexe masculin né le dix-neuf février, à deux heures du soir au village de Grudet, commune de Saint Symphorien, des mariés Dufournaud Alexis , cultivateur colon, âgé de trente ans , et Marie Pardonnaud, sans profession, agée de vingt-six ans , demeurant au village de Grudet, dite commune de Saint-Symphorien. [...] Father Mother Child			

Table 3.1 – Resulting networks using different models produced by one document of the examples detailed in §3.4.3: co-occurrence, unipartite and bipartite models. The first column shows the partial transcription of real documents. Colors represent annotations concerning the persons mentioned, their roles, and attributes. Underline refer to information related to the events and which can be encoded as document/event attributes. H: Husband, W: wife, T: Witness, M: Marriage, A_N: Associate, G: Guarantor, Ap: Approbator, C: Construction, F: Father, M: Mother, C: Child.

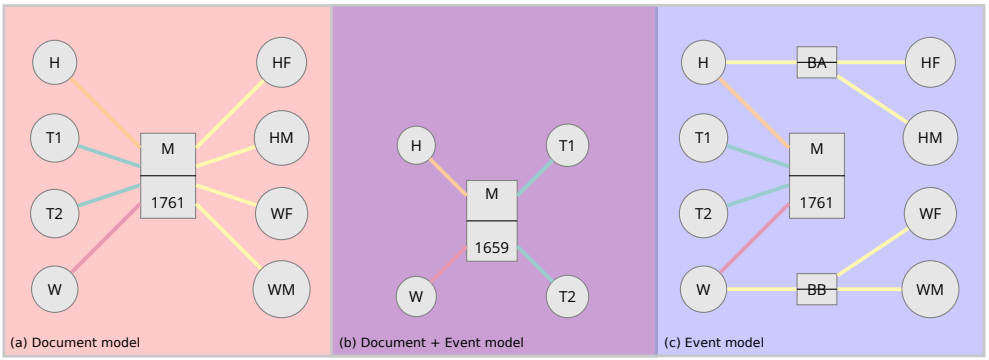


Figure 3.3 – bipartite multivariate dynamic network modeling for two cases of marriage acts of example #3. Some marriage acts mention the parents of the spouses, which is a relationship different than the marriage in itself. This case can be modeled using a document model (a) or an event model (c) by splitting the document into several different event nodes. The other case refers to documents that do not mention the parents (b) and in that case, the network represents both the documents and the events with the same model. M: Marriage, H: Husband, W: Wife, T: Witness, (H/W)(M/F): Husband/Wife Mother/Father. Yellow links refer to parenting mentions/relationships.

father and *wife's father* for example), thus turning the network into a model of the documents only, and not events. We show what would look like the resulting networks Figure 3.3for the two cases where marriage acts mention birth information and the case where only marriage-related information is present in the document.

3.7 . Conclusion

HSNA is a complex process that starts by collecting historical documents and ends with elaborating high-level sociological conclusions. Historians support their conclusions by modeling individuals' social relationships extracted from the documents and analyzing the resulting networks. We tried to shed light on this process by dividing it into 5 steps and describing recurrent pitfalls we encountered in our projects and collaborations. More importantly, we think this process should be done following the principles of *reality*, *traceability*, and *simplicity*, to avoid biasing the analysis, allowing to go back to the original source at any point of the workflow, and using models and methods simple and powerful enough for social scientists. Visual analytics software designed for HSNA should consider those principles to provide tools allowing to follow non-biased and reproducible analysis starting from the raw documents while supporting historians in going back and forth more easily between the annotation and analysis/visualization steps. We discussed the network modeling process in depth and claim that bipartite multivariate dynamic networks

satisfies those three core principles, letting historians both wrangle their data and characterize sociological phenomena using a common model and visual representation. Therefore, using this model VA interfaces could help social scientists manage and analyze their data starting at the data acquisition and annotations steps instead on focusing on the analysis only, while providing efficient representations of the data for analysis and exploration. We explore what could be such VA interfaces in the two next chapters.

4 - ComBiNet: Visual Query and Comparison of Bipartite Dynamic Multivariate Networks with Roles

In the previous chapter chapter 3, we decided with historians collaborators to model their historical documents into bipartite multivariate dynamic networks to follow HSNA analysis, as it satisfy *simplicity*, *reality* and *traceability* constraints. However, no visual tools currently exist based on this data model, to allow social scientists explore and analyze such network. In this chapter, we propose a VA interface aimed at exploring historical documents modeled as bipartite multivariate dynamic networks. We try to answer Q2 by analyzing tasks and questions historians have on their data and providing interactions mechanisms which would allow them to answer their sociological and historical questions.

This chapter is an updated version of an article currently submited to the journal Computer Graphics Forum, and a poster presented at the conference EuroVis 2022 [?]. It was a joint work with my collaborators Christophe Prieur and Jean-Daniel Fekete. I did the development of the interface, and participated in the discussions, evaluation, and writing of the paper.

4.1 . Context

Social historians and sociologists aim at retrieving and studying facts about a specific region and period of time that they focus on. Their work essentially relies on documents—such as marriage acts, census records, surveys, and business contracts—to gather information about the life of important actors that they explore in-depth, or to draw conclusions on social aspects of groups in the society of that period and place. Instead of drawing conclusions from their gathered knowledge and interpretations of the documents, a more systematic approach consists in constructing a social network from the documents and following a Social Network Analysis (SNA) approach [149]. For this, they need to encode their documents to extract the persons and any other useful information in the text and transfer it into a structured file or a database. Social scientists can then explore, validate, or refute their hypotheses by observing and analyzing the network structure and the connectivity patterns between the entities of the resulting network. They also want to visually explore their data to generate new insights and hypotheses.

Currently, social scientists often model their datasets as simple networks where the nodes are the persons mentioned in the documents (see chapter 3). Usually, Two persons are then connected together in the network when they appear in shared documents. This representation is easy to visualize and analyze but simplifies and

distorts the information by hiding the documents that witness the relationships between the persons. Thus, another approach consists in modeling the data as bipartite networks, where both the documents and the persons are represented as nodes and are connected together when a document mentions a given person [54, 120, 130].

In addition, historical documents include time and geospatial information corresponding to the date and location of the events they refer to. Documents often mention additional information on the persons, such as their sex, profession, and date of birth. These are often essential to understanding underlying social phenomena, as time, space, and social status play an important role in social dynamics. For these reasons, historical sources and the underlying social phenomena they refer to can be modeled well by *bipartite* with *roles*, *multivariate dynamic* networks. *Bipartite* means that both persons and documents (or events, that are often witnessed by physical documents) are modeled as typed nodes. *Multivariate* means that the nodes and links can carry additional attributes. *Dynamic* means that time is a mandatory attribute of documents. Furthermore, a link created between a person's node and a document's node (when the person is mentioned in the document), has an associated link type that models the *role* of the person in the document/event. Additionally, documents can optionally carry a geographical location. This model unifies several social network models and allows to model the historical sources with any transformation, simplification, or loss of information [23].

Several sophisticated tools exist to explore and analyze rich social networks. However, the majority of them either enforce too simplistic network models, such as Gephi [5] and NodeXL [133] or do not enforce any data model and lead to very complicated interfaces which are complicated to navigate for users like historians. Moreover, the majority of social network visual analytics tools provide limited interactions to query and explore richly encoded data, and historians often reach simple conclusions.

In this chapter, we present a visual analytic system to explore and analyze Bipartite Multivariate Dynamic Social Networks, with the aim of answering historical and sociological questions. We elaborated our tool based on four collaborations with social scientist colleagues. We first collected important questions they each had on their data and transcribed them from a network analysis perspective. The majority of the questions raised consisted in either finding specific patterns in the network or in comparing several subsets of the network, in terms of network measures, attribute distributions, and their overlaps.

we thus focus on three high-level tasks: exploration, queries, and comparison of this type of network. Users can explore the data using two layouts: a node-link bipartite view showing the sociological structure of the network, and a map layout based on the geolocation of documents. We designed and implemented a new visual graph query system that allows us to build both topological and attribute constraints, based respectively on a node-link interactive representation,

and dynamic widgets. For this, we rely on the Neo4j graph database [100] and its language *Cypher*. Most visualization systems offer dynamic queries to hide the complexity of query languages. However, using a rich data model, some queries are much easier to refine using scripting than dynamic queries. We implemented dynamic queries that also show the translated Cypher queries, and inversely, can translate textual queries into visual queries. With that interface, social scientists can start building their queries with simple widgets and, if needed, complement them by editing the query, alone or with the help of power users. On top of that, they can easily copy and paste the textual query to share the current state of the query and associated results with someone else or to start an analysis session from a previous result. ComBiNet also implements subgraph comparison techniques, allowing the comparison of networks, network-related measures, and attribute distributions between the entities returned by the queries. We validate the query and comparison system with a formative usability study and we demonstrate ComBiNet can be used to answer sociological questions by describing in depth several real-world use cases.

After the related work section, we describe our data model in detail using four use cases, and present our system ComBiNet, with the design of the visual query and comparison features. Finally, we present three use cases demonstrating the utility of our system, showing it can be used to explore complex historical data and allowing historians to answer several of their questions using queries and comparisons. Our contributions are:

- The design and implementation of a graph query system, synchronizing the visual representation of the query and the associated script;
- The design and implementation of visualization and interaction techniques aimed at comparing subgraphs, in terms of topology, attributes, time, and geographical location.
- A usability study and two real-world use cases demonstrate the utility of the system to answer socio-historical questions.

4.2 . Related Work

As we already discussed the related work on network modeling and social network visualization in chapter 2 we only discuss in this section visual graph querying, visual graph comparison, and provenance.

4.2.1 . Graphlet Analysis

One of the inspirations for this project came after participating in the 2020 VAST challenge [1] where we used graphlets to measure similarity between several

[1]This is a challenge organized in the context of the IEEE Visual Analytics Science and Technology (VAST) conference. The challenge consisted of a series of analyti-

networks [139].

Graphlets are small connected induced, non-isomorphic subgraphs composing any network. In an induced subgraph, two vertices linked in the original graph remain linked in the subgraph. For instance, if the original graph is a triangle we can only induce the simple edge or triangle subgraph (graphlet) The path of length 2 has all vertices of the original graph but misses an edge and is, therefore, not a possible graphlet. They were first introduced by Milo et al. [91] to explore the structural differences between biological networks, but they are now used in several disciplines involving networks such as sociology.

One of the aims of the VAST 2020 challenge was to compare several multivariate networks. However, by using graphlets we realized that 1) it was not very efficient to compare several networks in contrast to other measures, and 2) the interpretation of all graphlets patterns one finds in a network is not straightforward and can be complicated given the fact that one specific pattern can have various interpretations given the nodes involved and their positions in the network [68]. This is especially true that the number of potential graphlets grows exponentially if we increase the number of nodes considered (there are 6 graphlets of size 4 and 21 graphlets of size 5) and if we add complexity to the network model, for example by adding directed links or node and link types [118].

Instead of counting every graphlet occurrence and interpreting those with a sociological lens, social scientists are more interested in finding specific patterns to answer questions they ask themselves on the data.

4.2.2 . Visual Graph Querying

Several scripting languages, such as R [114] and Python [143], have been extended to support the exploration of social networks using specialized libraries such as igraph [27] and NetworkX [57]. However, social scientists are often challenged to use scripting languages and programming.

Finding and extracting a subgraph of interest in a bigger graph is an old problem in SNA. Constructing and querying a pattern from a graph requires knowledge of graph databases and query languages. To lower the complexity barrier, several visual graph query systems have been developed to allow analysts to rapidly build and refine their queries visually. GRAPHITE [20] and VERTIGo [28] allow specifying a graph query as a node-link diagram that the user creates interactively. Shadoan and Weaver [129] use a similar concept with hypergraphs to filter multidimensional data. Other systems, such as VIGOR [111] only visualize the query after it has been written using a scripting language. However, these visual systems are limited to topological queries, including constraints on the vertex and edge types; they do not support constraints related to general attributes and time associated

cal questions united under an overarching cyber threat scenario. We participated in the Mini-Challenge 1 which asked participants to identify a group of people that accidentally caused an internet outage. To identify this group, we were given a network profile and a large multi-variate social network to search in.

with vertices and edges.

4.2.3 . Visual Graph Comparison

Gleicher et al. [51] propose a taxonomy of visual comparison designs for complex objects. They claim any visual comparison system can be classified into one (or a mix) of the three following categories: juxtaposition, superposition, or explicit design. Yet, few visual systems support comparison tasks on social networks.

Andrews et al. [3] describe a technique to compare several graphs, using a combination of juxtaposition and superposition techniques. The two candidate graphs are shown side by side, along with a third view composed of a fusion graph highlighting both the shared nodes along with the non-shared nodes with different colors. Freire et al [44] describe the ManyNets system to compare many networks by using a table where each describes one graph and each column shows graph measures in terms of small visualizations, from simple bars to distributions, allowing the comparison of a large number of graphs. However, ManyNets does not visualize the networks per se (no layout shown), and do not take into account attributes, node types, or time. Hascoët and Dragicevic [60] describe a system to match and compare graphs using superposition, focusing on the topology, not taking into account attributes or time. Tovanich et al. [139] propose a visual analytics tool to compare multivariate, sometimes bipartite, dynamic graphs and find common structures. Yet, their tool does not handle roles and is designed for the specific task of matching a subgraph into a large graph.

4.2.4 . Provenance

Provenance in the context of Visual Analytics consists in the logging of the sequence of actions of users on an interactive visualization system during analysis sessions. Collecting provenance information has proven to benefit users by providing them action recovery (undo), and collaborative and reproducibility capabilities [115]. For example, VisTrails allows users to reproduce their visual analyses by providing an executable history graph of their actions, [17] while GraphTrail provides provenance tools to ease collaborative analysis [36]. Provenance can also be beneficial for visualization designers and researchers, as it gives them a tool to understand users' behaviors [7, 12] and evaluate/improve visualization systems [117]. All the reasons and concrete implementations of provenance are discussed in depth in Xu's survey [150].

4.3 . Task Analysis and Design Process

We designed the ComBiNet tool in collaboration with historians; all their historical documents data fitted well our bipartite multivariate dynamic network model. We first collected questions they had about their data and what they wanted to see in a visual interface. By analyzing the questions we leveraged tasks and requirements. We designed the interface from the requirements with continuous

Main Tasks	Subtasks	Views	Constraints
Bipartite Graph Exploration	T1.1 Overview of the network	V1	A node-link representation is expected. The geolocation of events has to be done according to the historical period.
	T1.2 Overview of nodes attribute values and distributions	V1,V2,V4	
	T1.3 Show the persons' roles in the documents they appear in	V1	
	T1.4 Show the location of the different documents	V2	
	T1.5 Show the time of the documents	V1,V2,V4	
Apply filters to isolate subgraphs	T2.1 Filter on topological patterns	V6,V8	Constraints must be easy to set and visual.
	T2.2 Filter on attribute values	V7,V8	
	T2.3 Show the provenance of filters	V9	
	T2.4 Show the subgroups alone or in network's context	V1,V2	
Compare several subgroups	T3.1 Show the shared and exclusive entities	V1/V2	
	T3.2 Compare the node attribute distributions	V4	
	T3.3 Compare the subgraph measures	V3	

Table 4.1 – Tasks to support during exploration, according to our expert collaborators, split into 3 main high level tasks.

discussions with our collaborators. We showed them visual prototypes during the development phase to get feedback iteratively.

4.3.1 . Use Cases

We elaborated this interface from the collaborations with historians we described in §3.4.3. These collaborations involved regular meetings and multiple interviews over two years. All these datasets are textual corpora constituted of historical documents mentioning people with complex relationships. They are well modeled by bipartite multivariate dynamic network. We give more details about the datasets of these collaborations in this section and we also list our collaborators' main questions and the graph queries extracting the information to start answering them. The full answers involve visualizations of the query results and attribute summaries that we describe in the next section. We list the most important questions our collaborators shared with us on their respective datasets. We categorized those according to four dimensions: global (G)/local (L) (do they want to categorize a group of nodes or retrieve specific persons/documents), if the question can be answered using the topology (T), and/or the attributes (A), and finally if a comparison (C) using several filters is needed or not (N).

1. Analysis of the social dynamics from **construction contracts in Italy in the 18th century (141 documents, 272 persons) [25].** The corpus is made of contracts (manuscript documents) for different types of constructions in the Piedmont area in Italy. People are mentioned in three different roles: *Associates*, who participate in the construction; *Guarantors*, who bring financial guarantees; and *Approvers*, who vouch for the guarantors. Along

with the time and location of the construction site, documents have a construction type (military, religious, and civil), work type (big work, small work, reparation, transportation, etc.), and material (wood, stone, metal). People also have an origin attribute (the place they come from), manually extracted from the original documents.

Question 1 Do approvers act as bridges compared to associates and guarantors? (G, T, C)

Query 1.1 Request all approvers occurrences

Query 1.2 Request all associates and guarantors occurrences

Question 2 What are the differences between Turin (Torino) and Torino close area according to the contracts? (G, AT, C)

Query 2.1 Request all documents located in Torino, with the persons mentioned

Query 2.2 Request all documents located in the Torino area, with the persons mentioned

Question 3 Who are the persons of the extended Zo family (G, AT, N)

Query 3.1 Request all the persons of the Zo family and their N+2 ego network

Question 4 Compare the Menafoglio and Zo families in terms of contracts and activities (G, AT, C)

Query 4.1 Request all the persons of the Menafoglio family and the documents that mention them

Query 4.2 Request all the persons of the Zo family and the documents that mention them

Question 5 Who are the persons having the 3 roles? (G, AT, N)

Query 5.1 Select persons with associate, guarantor, and approbator roles in 3 different documents

Question 6 Are there people mutually guarantors to each other in different contracts? (G, AT, N)

Query 6.1 Select pairs of people connected each to the two same document, with a guarantor role and any other role

2. Analysis of migrations from the **genealogy of a french family between the 17^{th}–20^{th} centuries (2053 events, 957 persons from a private source).** The corpus is made of family trees referring to several document/event types: birth and death certificates, marriage acts, military mobilization, and census report. The roles are different for each event type and consist of *children, father, mother* for the birth events, *deceased* for the death event, *spouse* and *witnesses* for the marriages, and *family member*s for the census events.

Question 7 Overview of the trajectory of life for an individual (birth, living, marriage, death) (L, A, N)

Query 7.1 Select one person and all her/his documents (to use the mentioned places)

Question 8 Overview of the trajectory of life for a family (L, A, N)

Query 8.1 Select birth certificates with the child, parents, and birthplace

Question 9 What are the main migrations? (G, A, N)

Query 9.1 Select persons with a geolocated birth and death certificate

Question 10 Is there differences between migrations in the 18th and 19th centuries? (G, A, C)

Query 10.1 Select persons with a geolocated birth and death certificate from the 18th century

Query 10.2 Select persons with a geolocated birth and death certificate from the 19th century

Question 11 In the Haute-Vienne and Cote d'Armor administrative areas, are there cycles in living places every 10/20 years? (G, A, N)

Query 11.1 Select persons with their census reports located in Cote d'Armor and Haute-Vienne

Question 12 In the 19th century, was there an overall decrease in the social status and professions of persons in the dataset? (G, A, C)

Query 12.1 Select persons in the first half of the 19th century with a profession mentioned

Query 12.2 Select persons in the second half of the 19th century with a profession mentioned

3. Analysis of migrations from Spain to Argentina through the **marriage acts at Buenos Aires in the 17–19th centuries (1396 acts, 6731 persons) [96].** The corpus is made of acts that mention the spouses and the witnesses of the wedding, which are the roles modeled by the links. The origin, date of birth, and parents' names are specified for both spouses.

   Question 13 How are spouses and witnesses linked in their family network? (G, T, N)

   Query 13.1 Select marriages with spouses and witnesses, where the spouse and witnesses have the same parents

   Query 13.2 Select marriages with spouses and witnesses, where the spouse and witnesses have the same grandparents

   Question 14 Who are the persons with 2 marriages with a long delay? (L, A, N)

   Query 14.1 Select persons in 2 marriages as husband or wife. Put a constraint on the difference of time in the marriages

   Question 15 Where are the persons marrying in Buenos Aires coming from? (G, A, N)

   Query 15.1 Select persons with a birth certificate located not in Buenos Aires

4. Socio-political analysis of **migration of ethnic Germans from communist Romania to West Germany in the 20th century (ongoing work) [32].** The corpus is made of administrative forms that mention persons requesting to migrate, along with the persons they want to join, and the administrative persons of the ministry in charge of the forms (3 roles). The family members of the aspiring migrant are also mentioned in the forms, with their respective dates of birth.

Question 16 What member of their family do emigrants usually join? (G, AT, N)

Query 16.1 Select all migration documents with the emigrant and the person they are joining

Question 17 What price does the emigrant have to pay, given their socio-economic profiles? (G, A, C)

Query 17.1 Select people who are mentioned in a budget and a migration document

4.3.2 . Tasks Analysis

Most of the questions we collected from our collaborators could be answered by isolating a subgroup of entities and analyzing them in the context of the whole network, or by comparing two subgraphs, in terms of their entities, structure, and attribute distributions. From discussions with our collaborators and the analysis of their questions on their data, we elaborated a list of requirements for the visual interface, split into three main parts: 1) Exploration of the data, 2) Queries, and 3) Comparisons. The elaboration of the tasks was an iterative process, as we showed the interface to our collaborators several times in the development phase to get feedback. The tasks are described here and summarized in Table 4.1:

1. **Exploration of bipartite multivariate dynamic network**. The visual interface must allow exploration of this specific type of network, using every aspect of the data, i.e. its topology (T1.1), node attributes (T1.2), roles (T1.3), geolocation of the documents/events (T1.4) and time (T1.5). Common interactions such as selection and zooming are also needed for the exploration.

2. **Applying filters**. To answer their questions, users need to be able to apply filters to the data, to isolate specific groups of entities having specific behaviors or characteristics. To answer the diversity of questions, they should be able to put constraints on every aspect of the data, i.e. the topology, the roles (T2.1), and the attributes (including time and geolocation) (T2.2). Access to provenance information can also help them in their query construction, by going to previous states and exploring different paths more easily (T2.3). Once they are satisfied with their query, they want to explore the results, usually in the context of the whole network (T2.4).

3. **Comparison of several subgraphs.** Users should be able to compare several subgraphs isolated after applying filters, to see the similarities and differences between groups of entities of interest. The system should be able to easily see the common and shared entities of the two subgraphs (T3.1), their respective place in the network, their structural differences (T3.2), and their different attribute distributions (T3.3).

	Bipartite	Node Attributes	Links Attributes	Dynamic	Geolocated
Jigsaw	✓	Only some	✗	✓	✓
Puck	✓	✗	✗	✓	✗
ComBiNet	✓	✓	Encode roles	✓	✓

Table 4.2 – Comparison of the data model of several VA systems aimed at exploring bipartite social networks.

4.4 . The ComBiNet System

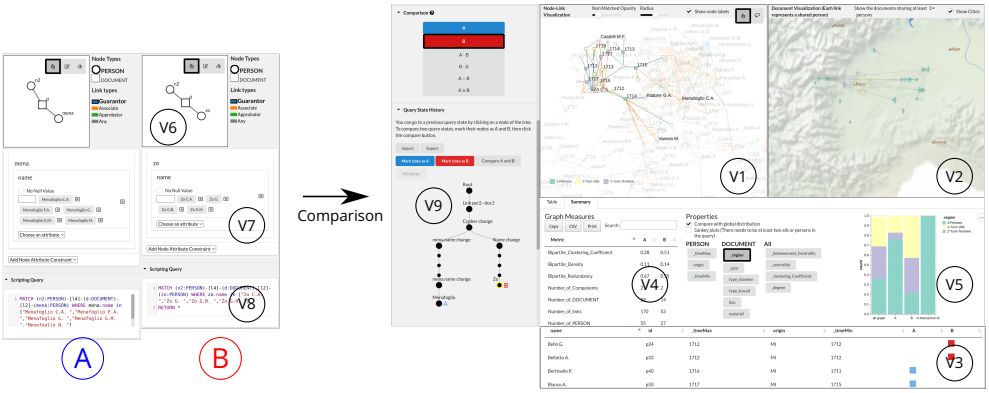


Figure 4.1 – The ComBiNet system used to compare two subgroups of a social network of contracts from [25], extracted with dynamic visual queries. (A) and (B) show the two visual queries created by the user in the query panel using an interactive node-link diagram editor (V6), dynamic query widgets (V7), and the equivalent Cypher script (V8). The right part shows ComBiNet's global interface in *comparison* mode: (V1) Network visualization panel, (V2) Map of the geolocalized nodes, (V3) Table of persons, (V4) Graph measures comparison, (V5) Attribute distribution plots, and (V9) Provenance tree. The two visual queries on the left, translated into Cypher queries below, select the "Menafoglio" family on the left, and the "Zo" family on the right, along with their construction contracts and close collaborators.

ComBiNet is designed to visualize, explore, and analyze social networks encoded as bipartite multivariate dynamic network. Some other systems exist to explore bipartite social networks such as Jigsaw and Puck, but do not encode every aspects of historical documents historians are interested in. See Table 4.2 for a comparison of their data model compared to ComBiNet.

When started, it dynamically collects the node types, roles, sub-types, and attributes when reading the network from the database. ComBiNet is constituted of four main panels, split into different views as shown inFigure 4.1: the query and comparison panel, the graph visualization panel, the map visualization panel, and

the query results panel.

4.4.1 . Visualizations

ComBiNet presents a social network with multiple visualizations highlighting different aspects of the data. The visualizations are linked when it makes sense so that interactions such as selection done on one propagate to other panels.

V1: Bipartite Node-Link Diagram The bipartite node-link visualization panel shows the network using the DrL force layout from igraph [27] with overlap removal using D3 [13]. Node-link representations are very common in social sciences [5, 98, 133] and were a specific request from our collaborators. In the context of our bipartite model, the persons are represented as circles and the documents/events as squares, while the roles are encoded as link colors. A link models the mention of a person in a document. This view provides an overview of the data by showing the structure of the network (T1.1) and the roles of the persons in their different documents (T1.2). Attribute values can be overlayed on the nodes using colors when users select an attribute. It allows detecting patterns relative to attributes, in the context of the topology of the network (T1.2, T1.4, T1.5). For example, Figure 4.2 shows the construction dataset of #1 where the user selected the *year* attribute, coloring the documents nodes with their year in the node-link diagram (left). The view also provides pan & zoom and selection interactions for effective navigation.

V2: Map View The map visualization panel on the right shows an event-centric view, displaying only the geolocalized event nodes on a map. By default, only event nodes are shown, but users can select a threshold to show links between nodes when they share at least a given number of persons in their mentions. Persons are not directly shown in this view as they do not have a unique location. This map view presents a transformation of the bipartite graph, focused on the geospatial information that is very important to social scientists (T1.3).

As we collaborate with historians who study different periods, we cannot use modern map backgrounds such as the default one provided by OpenStreetMap or Google Maps since many features are anachronistic (e.g., roads, administrative areas, borders). We, therefore, provide a map background with only these non-administrative features: elevation, lakes, rivers, and types of environment. We also show the most important cities as most of them existed in the past and provide landmarks. The map uses Natural Earth tiles and vector data [99].

The two views are coordinated: selecting/hovering an event node in the graph view highlights it on the map and vice versa, while hovering a person node highlights all its corresponding documents on the map, rapidly showing the person's events' locations.

V3: Entities Tables All the persons and the documents of the loaded dataset are listed in two separate tables, showing the attributes of the entities. This way, users can order the entities according to any attribute they want (T1.2). The tables are linked to the visualizations, meaning that selecting a row highlights

the respective entity in the visualizations, and vice-versa. Tables in social network visualization systems have been proven to be efficient and useful for social scientists when exploring their data [9]. It allows them to link the visualization to the network entities more easily, and dive deeper into one entity's attribute values after selecting it in the network. It also makes ranking entities according to various criteria easier and more straightforward.

V4: Graph Measures The Graph Measures view shows measures related to the network and gives insights into its structure to users (T1.1). We report simple measures like the number of persons, documents, links, and components, and more sophisticated bipartite network measures asked by our users, that they can report for their analysis: the bipartite centrality, bipartite clustering coefficient, and bipartite redundancy. **explain measures** These measures are updated in real-time when filters and comparisons are applied.

V5: Attributes View All the attributes in the network are shown as buttons in the bottom right of the interface, sorted by their associated node type (person, document, and both). They can be quickly visualized by hovering over the button, producing two effects: it colors all the nodes on the two views according to their attribute values, and it shows a plot of the distribution of the selected attribute, as shown in Figure 4.2. By clicking on the button, the visual encoding and distribution remain selected. This interaction is inspired by the x-ray technique of the Vizster system [61]. Users can follow a first exploration of their data by visually detecting correlations between attribute values and some groups of persons or between attribute values and some specific areas in the map view (T1.2, T1.4, T1.5).

4.4.2 . Query Panel

The query panel allows to rapidly build queries visually, with topological and attribute constraints. The visualization of the query is synchronized with the Cypher query sent to the database. Modifying one representation will update the other, allowing users to build a query visually and refine it in Cypher when appropriate. Experts users who know the Cypher language can also start to construct their query textually and modify it visually later on. In this section, we describe all the features and interactions allowing ComBiNet to build a query and illustrate them with questions 2 and 6 of the use case #1. Our collaborator wants to *find the persons who are mutually Guarantor to each other in separate contracts (6)* and to know *how Torino and Torino's surroundings differ according to their contracts?*.

V6: Node-Link Dynamic Query

The interactive node-link diagram allows building a subgraph query graphically, which represents a topological constraint (T2.1). The query subgraph is built and edited interactively. At each modification, the subgraph is converted into a Cypher query, run in the database, and all its matches are returned and highlighted in the main visualizations. Three modes of interaction are available through the top-right menu: *selection, addition*, and *deletion*. The *selection* mode allows to drag

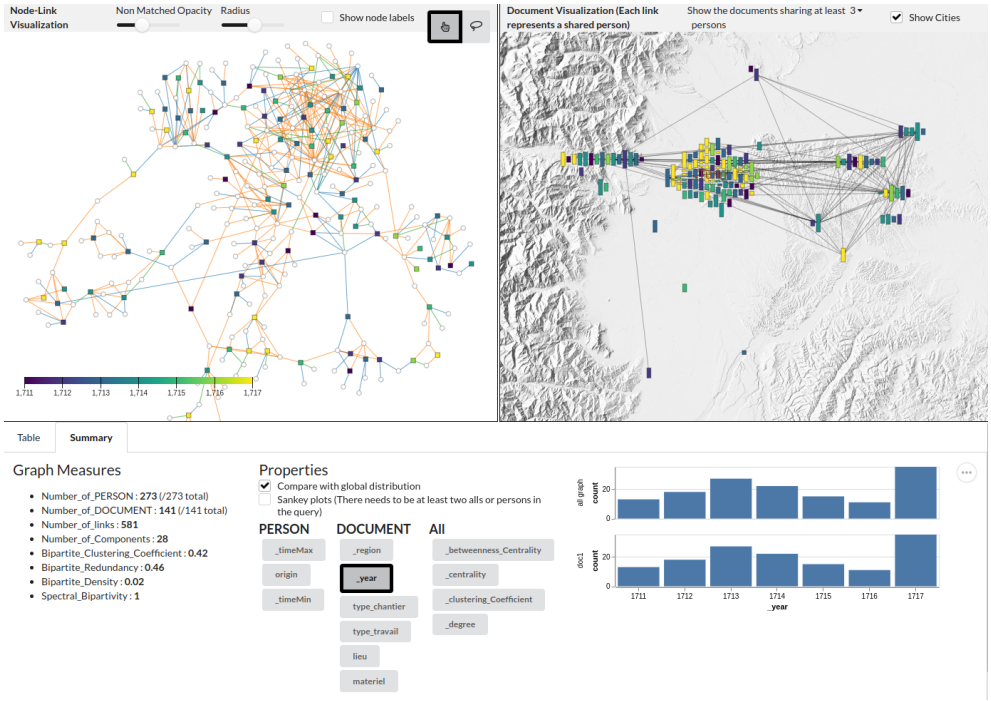


Figure 4.2 – ComBiNet interface with the dataset of collaboration #1. The user selected the year attribute, showing the distribution of document years with a histogram (bottom right), and coloring the documents node on the bipartite view (left) and map view simultaneously (right).

the nodes in the panel, while the *addition* and *deletion* modes allow the following actions:

Node Creation: In *addition* mode, clicking on an empty area creates a new node. The node will be of the selected type from the legend on the right (Person, Document, or Any).

Node Deletion: In *deletion* mode, clicking on a node deletes it and its links.

Change Node type: In *selection* mode, clicking on a node opens a menu allowing to change its type.

Link Creation: In *addition* mode, clicking on a node and dragging the mouse to another node will connect the two with a link. Its type (color) will be the link type selected on the legend.

Link Deletion: In *deletion* mode, clicking on a link deletes it.

Change link type: In *selection* mode, clicking on a link opens a menu to change its type.

Users build concrete subgraphs with the same representation as in the bipartite graph view: a visual query is a graph template. Each role (link type) is rendered using a color (Figure 4.3 left). We can also create untyped links using the *Any* value, which will be matched by all the existing link types (Figure 4.3 left). We also allow creating links that can be matched by several selected link types in the graph, by checking several possible types for one link. These links are represented by a dashed line with the colors of the possible types (Figure 4.3 middle right). Several links with different types can also be created among two nodes to query a person with more than one role in the same event (Figure 4.3 right). When a node or link is created in the query, it is given an identifier starting with *per* for a person, *doc* for a document, *link* for a link, followed by a number. These identifiers are used in the attribute constraint panels and the textual query and can be changed through their textual representations.

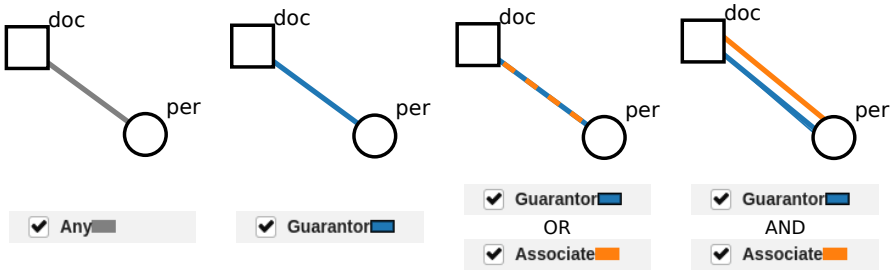
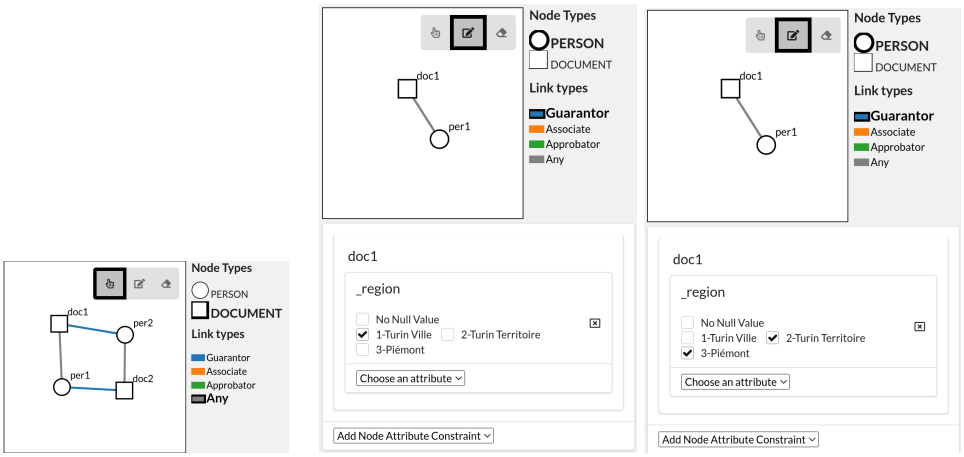


Figure 4.3 – All link creation possibilities: Any link type (left), one selected link type, here guarantor (middle left), the union of several link types (middle right), several links with different types (right)

To find persons who are mutually guarantors in our collaboration #1, we first create one person and two documents using the addition mode and by clicking on the canvas. We then link the person node to the first document with a link that is not typed (Figure 4.3 left), and link it to the second document with a Guarantor link (Figure 4.3 middle left). We then create a second person node and link it to the two documents with opposite link types. The resulting visual query is presented in Figure 4.4 (a). To answer the second question, we can simply start to request all the links in the graph, no matter the type, as shown in Figure 4.4 (b). The database will then return all the links in the graph with their attached nodes.

V7: Attribute Constraint Widgets Users can also add attribute constraints (T2.2) on the created nodes with the help of interactive widgets. An input button is created for each node and link identifier from the node-link query panel. It allows to create a dynamic query widget for any of its attributes. The widget design will vary according to the three possible attribute types: numeric, categorical, or nominal, as in the original dynamic queries [131]:

1. **Numeric constraints** are modeled as range sliders, allowing to select a lower and upper bound to the filter.



(a) (b)

Figure 4.4 – Visual queries created to answer questions 2 and 6 of our collaboration #1. (a) The visual query retrieves individuals who are mutually guarantor to each other in separate construction contracts. (b) The two visual queries retrieve the documents—along with the signatories—of *Turin* (left) and of *Turin Territoire* (right)

2. **Categorical constraints** are modeled as a set of checkboxes. Each possible value has a corresponding checkbox.

3. **Nominal constraints** are modeled as text input, where the user can write any desired value. All the possible values are shown at the same time and filtered as the user writes.

For the categorical and nominal widgets, selecting several values will correspond to the union of the filters. The three widget types are shown in Figure 4.5.

To answer our collaborator's second question (*how do Torino and Torino's surroundings differ according to their contracts?*), we first want to filter the documents which are located in Torino (*Turin* in French). For this, we start by selecting the whole dataset by linking a person and document node with *any* link. Then, we select the id *doc1* of the document of our visual node-link query, and the *region* attribute. It will initialize a categorical widget including all the values found in the dataset for this attribute with associated checkboxes. We check the region of interest "*1-Turin Ville*" to select all the documents from this region. The first widget of Figure 4.5 illustrates the created constraint. To select the documents of Torino's surroundings, we can simply uncheck the "*1-Turin Ville*" value for the *region* attribute and check the two other values "*2-Turin Territoire*" and "*3-Piemont*" which are areas corresponding to the surroundings of Torino. Both queries are represented in Figure 4.4 (b).

V8: Cypher Editor Users can build or modify a query using the Cypher query

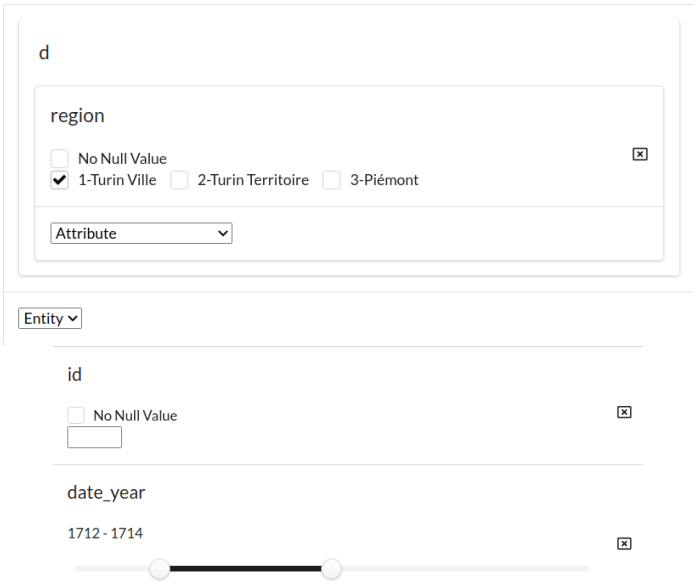
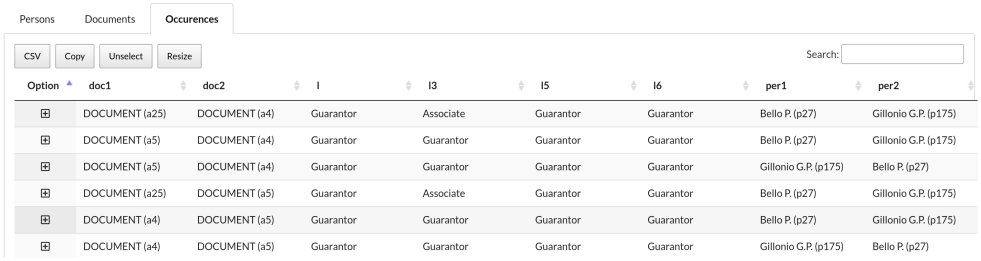


Figure 4.5 – Widget designs for the different attribute types: checkboxes for categorical attributes (top), text input for nominal attributes (middle), and a double slider for numerical attributes (bottom). The categorical attribute example shows the options input letting users create new constraints for other attributes and other nodes.

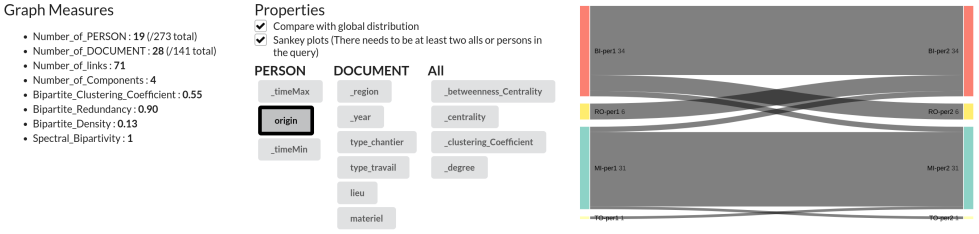
language, with the Cypher text editor. This allows users to start creating a query visually and refining it by text for complex constraints which can not be represented by a visual form easily. The editor supports autocompletion to e.g., help to discover and spell the attribute names. The visual and textual representations are synchronized, meaning that changing one will update the other and update the results in the visualizations.

Query Results Each modification of the query, whether from the node-link dynamic query, the widgets, or the Cypher text boxes, update the two visualization panels (V1, V2), the entities tables (V3), the graph measures view (V5), and the attribute plots (V6). The nodes and links that do not match (are not retrieved by the query) are grayed out in V1 and V2 and are removed from the persons and documents tables (V3). A third table shows every occurrence found of the created pattern that we call the occurrence table. The occurrence table for question 1 of collaboration #1 is shown in Figure 4.6 (a). It tells us that the pattern has been found 36 times. Users can switch between the three tables in the table view using the tabs. The graph measures are computed on the new graph formed by the union of all patterns found and updated on the graph measures view (V5). Figure 4.6 (b) (left) shows the user the different graph measures of the subgraph induced by the patterns found. Since some measures can be long to compute, the values are computed iteratively in the backend and shown progressively [41] to

avoid blocking the interface. The distribution plots in the attributes view (V6) are updated, showing the values of the entities of the latest constructed query, next to the global distributions.



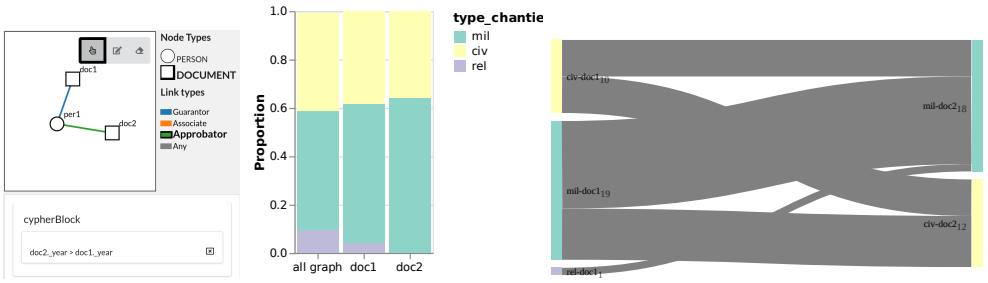
(a)



(b)

Figure 4.6 – Results of question 1 of collaboration #1: (a) shows a subset of the table view with every occurrence of the pattern found. (b) shows the summary panel, with the graph measures and the attributes view with the *origin* attribute selected and the Sankey option checked. It allows us to see the attribute distribution of the persons included in the pattern and see if there is a relationship between persons who are mutually guarantors and their origin.

Attributes Visualization When users select an attribute in the attributes view (V5), its distribution is visualized for the queried entities and the whole network. However, these plots show the aggregated values and we lose the potential value transitions between the query nodes. For example, Figure 4.7 shows a query to list the persons with the role of "approbator" (green) in a contract after being a "guarantor" (blue) in another contract (using a time constraint). We may want to see if the locations or types of the two contracts are the same or if they change, case by case. Unfortunately, we lose this information with the aggregated plots. By checking the "Sankey" option on top of the distribution visualization, the plots are transformed into Sankey diagrams, giving information on how the attribute values relate between the nodes (person or event) of the same query. A Sankey diagram showing the attribute distributions is particularly useful for queries where the nodes have intrinsic time relationships, such as birth certificates, marriage, or death certificates where we know the order in which these events occurred. It is



(a) (b) (c)

Figure 4.7 – Two ways of showing the distribution of "type chantier" (type of works), a categorical attribute with three possible values "*reli*gious", "*mil*itary", and "*civ*ilian". (a) A query matching the contracts made by the same person (*per1*) as an "approbator" (green link to *doc2*) after being a "guarantor" (blue link to *doc1*) using the constraint (`doc2._year > doc1._year`). (b) Stacked bar chart for the matches, the earlier contract (*doc1*), the older contract (*doc2*), and (c) Sankey diagram with the early values on the left and the last on the right. The Sankey diagram reveals the value changes between the two documents: the guarantor who worked initially on religious work switched to military work.

also useful for queries with user-defined time order constraints as in Figure 4.7. The graph measures and attribute visualization view for the results of question 1 of collaboration #1 are shown in Figure 4.6. The Sankey view of the *origin* attribute shows that mutual guarantors come from 4 regions only and that usually, people have mutual guarantor relationships only with persons of the same origin. This is especially true for persons from *Milano*, and with some reciprocal links between persons from *Bioglio* and the *Comune di Ro*.

V9: Provenance Tree Each change in the query panel is saved with the computed results so that the history of the query construction can be shown in the form of a provenance tree (T2.4), managed using the Trrack library [29]. Each node of the tree represents a query change, with a description label like "New Link". It allows to rapidly visualize the succession of filters applied with their refinements. At any moment, users can click on a tree node to go back to the previous state; allowing them to navigate in the exploration states. Hovering over a node shows a tooltip with the query panel associated with the selected query state. It let users rapidly see what query is associated with each node of the tree If a new change is made on the query from a previous state, a new branch is created on the tree, allowing to revisit and refine explorations. Figure 4.8 shows the provenance tree made to answer question 2, split into 2 branches, with the tooltip showing one of the node query state.

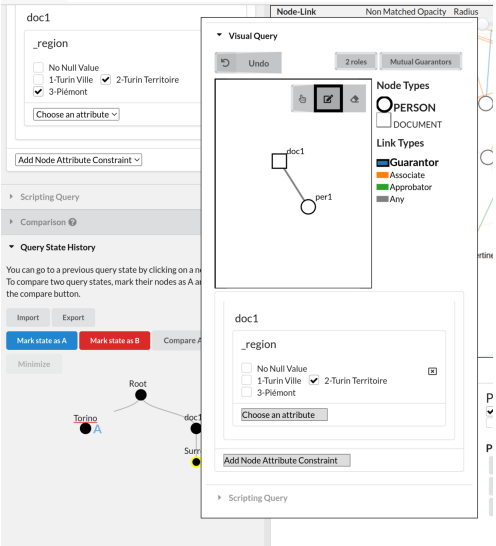


Figure 4.8 – Provenance tree to answer question 2 of collaboration #1: left branch leads to Torino documents (the node is labeled as A) while right branch leads to surrounding documents (the node is labeled as B). The user hovers over one node, revealing a tooltip that shows the visualization of the node's query..

4.4.3 . Comparison

In addition to comparing the results of a query to the whole graph, ComBiNet allows comparing the results of two queries. Users can select two query states in the provenance tree and mark them either as "A" or "B". Clicking on the button "Compare State A and B" compares them. The interface changes to *comparison mode*. Several buttons appear on top of the provenance tree: A, B, $A-B$, $B-A$, $A \cap B$, and $A \cup B$ for exploring the combinations of the two results of A and B in the two visualizations panels.

To answer several of the questions raised by our collaborators, we need to compare two subsets of the network.

For the second example from Table 4.1, we want to compare the works in Torino with the ones in Torino surrounding. Since we previously constructed the query returning all the contracts from *Turin* with the mentioned people, we can return to this point in the provenance tree, and change the constraint of the *region* attribute from *1-Turin Ville* to *2-Turin Territoire* and *3-Piemont* using the checkbox to get the two queries we want to compare. They are shown in Figure 4.4. The user can then rename the provenance tree nodes with explicit names such as "Torino" and "Surroundings", and mark them as A and B using the appropriate buttons. Clicking on the "Compare State A and B" will make the interface compare the two query results.

Topological Comparison In visualization mode, users can rapidly switch between

the visual filters of (A) and (B) by hovering over their respective buttons on the comparison menu and thus compare the structure of the two resulting subgraphs (T3.1). Similarly, different boolean comparison operations are available by hovering their respective buttons (Figure 4.1-C), such as the intersection, union, and differences between the two filters. Moreover, the summary tab (top of Figure 4.1-D) allows comparing the different graph measures of the two subgraphs by showing them side by side (T3.3). Comparing these measures, such as the number of matched documents or the densities, is crucial for SNA.

Metric	A	B
Bipartite_Clustering_Coefficient	0.52	0.57
Bipartite_Density	0.04	0.03
Bipartite_Redundancy	0.45	0.38
Number_of_Components	13	25
Number_of_DOCUMENT	42	46
Number_of_links	153	155
Number_of_PERSON	99	119
Spectral_Bipartivity	1	1

The

Figure 4.9 – Comparison table of the graph measures the query filters (A) and (B)

Check la comparaison

comparaison de torino et territoire sur le graphe

Attribute-Based Comparison The comparison of one or several attribute distributions between (A) and (B) is also useful for answering the historical questions of our users. In the attribute view (V5) of the results panel, hovering or clicking on an attribute name will show the distribution of this attribute in four contexts: the nodes of the whole graph, the queries (A), (B), and the currently selected Boolean operator (e.g., intersection or union) if one is selected. This allows users to compare attribute distributions between several subsets of interest (T3.2). For example, we can compare the attributes between the contracts of Torino and the ones of its surroundings. We can also compare the persons who worked in Torino, in Torino's close territory, and in both areas, by selecting the intersection operator. Figure 4.10 illustrates the comparison charts for different attributes. We can see that the types of construction sites differ between the two regions: the city of Torino clearly has a lot of military sites compared to the surroundings of Torino, which has almost none. This is the opposite for the number of religious sites, which are almost all localized in the surroundings of Torino. If we now look at the year distribution of the contracts, we can see a difference in the distributions. The

years of Torinos's construction contracts were steady between 1711 and 1717 with a little spike in 1713, while the constructions were more scarce in the surroundings before 1716. We can see a big spike in construction in 1717. This is interesting to our users, as it shows the dynamic of the construction in the area: the center of the city started to be constructed before other constructions arose in the surroundings.

We can also compare the profile of persons who collaborated at Torino and Torino surroundings by selecting the intersection of those two queries. One of the questions the historian had (question 2 of Table 4.1) was to know if those persons were a group with specific attributes and characteristics, or were inseparable from other persons working in the two areas. If we look at the betweenness centrality, on average, the values are higher for this group of people, meaning that the persons who work on the construction site at Torino and Torino's territory are clearly two distinct groups, and the persons collaborating in the two areas act as bridges between these groups. This visual demonstration was convincing and revealing for our users.

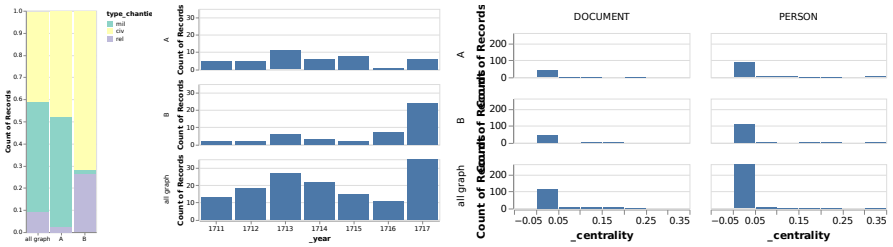


Figure 4.10 – Distribution of the type of constructions, the years, and the centrality for the documents and signatories of Torino (A), Torino surroundings (B), and the whole graph. (top).

4.4.4 . Implementation

ComBiNet is made of three components: a web visual interface, a python server, and a Neo4j graph database instance. The client interface is written in JavaScript using D3 [13], Vega [125], and the Trrrack library [29]. The python server is written in Flask and interacts with the Neo4j instance for query processing before sending the results to the frontend. We implemented our Cypher parser with the ANTLR parser generator [107]. **Talk about AST and implementaion**

4.5 . Use Cases

In this section, we describe how our system has been able to specifically answer questions from two of our collaborations. the tool was mostly operated by the developers working side by side with the collaborators to test the expressiveness of the queries and the value of the results visualizations. The tool was refined as needed along the way.

4.5.1 . Construction sites in Piedmont (#1)

One of the main questions of our collaborator was to compare two families which he knew played a big role in the structure of the network: the *Menafoglio* and *Zo* families (question 4 in Table 4.1). Specifically, he was interested in knowing if there were differences in specialization in type of contracts and area of work for the core members of these families, and to what extent the two families were collaborating. Moreover, he was very interested in characterizing the group of people collaborating with both families.

To answer those questions, we first selected the core members of the *Menafoglio family*, by checking the people known by the historian, and their close neighbors. Looking at the bipartite view (see Figure 1 of the supplementary material), we can see that the group is pretty dense with people collaborating a lot between them. Looking at the map, we can clearly see that the family has been mostly active in Piedmont outside of Torino and Torino's close territory. We also have a first view of the attribute distribution of the persons in the group and their contracts.

We then do the same query for the Zo family. We keep the same topological filter and replace the name filters with the core members of the Zo family known by the historian. We see on the graph view (Figure 2 of the supplementary material) that the group is smaller and is in a different area in the graph. The map enriched with a selection of the *region* attribute shows that, contrary to the Menafoglio, the Zo family has been more active in Turin and around.

The two groups can be compared using the *comparison mode* by selecting the two queries in the provenance tree. This opens the comparison menu to quickly navigate between the visual selection of (A), (B), and the set $A \cap B$ that interests our collaborator. The table showing the graph measures of the two subsets confirms what is shown visually: the Menafoglio group is more populated but less dense than the Zo family.

Our user is then interested in comparing the distribution of several attributes between the two groups. We can clearly see in Figure 4.11 (middle) that the Menafoglio family is more specialized in military sites, while the Zo family is doing more civil construction. This is confirmed by the "material" distribution that shows that the contracts of the Menafoglio are often using stones, whereas it is never the case for Zo contracts. Finally, the persons collaborating in the two groups have a betweenness centrality higher on average. This makes sense as they act as bridges linking the two families.

4.5.2 . French Genealogy (#2)

We describe how ComBiNet allowed us to answer an important question of the use case #2: to detect the largest migrations across several generations, in which areas, and at what time they occurred (question 7 in Table 4.1). The map view shows at a glance (Figure 3 in the supplementary material) that the majority of events have taken place in three specific regions west, mid-north, and mid-south.

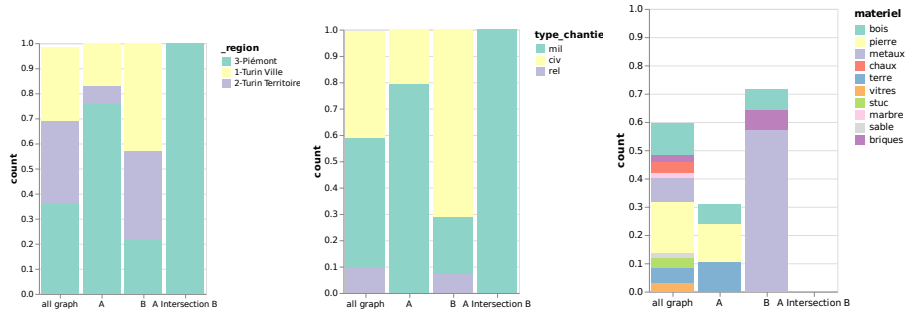
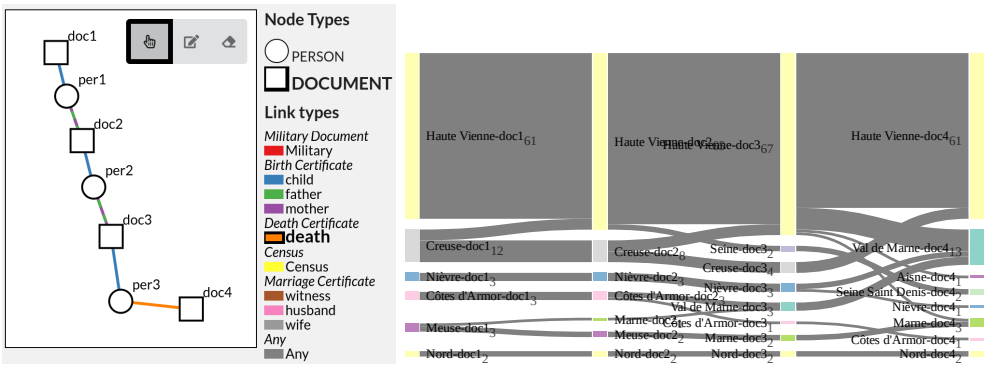


Figure 4.11 – Attributes distributions plots between the whole graph, the *Menafoglio* family (A), the *Zo* family (B), and $A \cap B$, for the *region,type_chantier, material type.*

To find patterns of migrations within families, we first make a query representing a simple family by linking a person node to a birth event, connected to the parents using a link of *father* or *mother* type. We repeat the process to the new parent node to add another generation. Finally, we connect the latest generation child with a death event, to have another date and location to compare to (see Figure 4.12a). This query returns every person with their parents and grandparents, along with their respective birth and death data for the latest person. We also create a constraint on the *department* attribute on the documents to only retrieve the events that have a non-null associated location. This request returns a subgraph of 64 persons and 88 documents. The user can now select the *department* attribute to create a Sankey diagram that shows the change of departments across the different generations of the families. Figure 4.12b shows that the majority of families are from *Haute-Vienne* (which can easily be confirmed by checking the map), and do not move much across generations. Our collaborator however detected interesting patterns of people moving from the department *Creuse* to *Haute-Vienne* across two generations. She refined the query by adding an attribute filter on this specific department using a widget. The table view then showed her who these migrants were and when it occurred. The bipartite visualization panel allowed exploring more in-depth this specific group of people.

Afterward, we answered question 8 (Table 4.1), to compare the migrations between the 18th and 19th centuries. She thought people started moving in the 19th century and wanted to confirm it. To answer this, we first created a query to retrieve the people with birth and death certificates from a specified department. We then applied a time filter on the death certificate node, first for the 18th century and then the 19th century, compared the two query results using the comparison mode, and looked side by side the Sankey graphs related to *departments* (Figure 4.13). We can clearly see that people do not move at all in the 18th century, while in the 19th century even if the majority of people stay in the same place from their birth to their death, more than half moved.



(a) Visual query to find all 3-generation families

(b) Sankey diagram showing the birth and death places of people across generations

Figure 4.12 – Migrations across departments over three generations

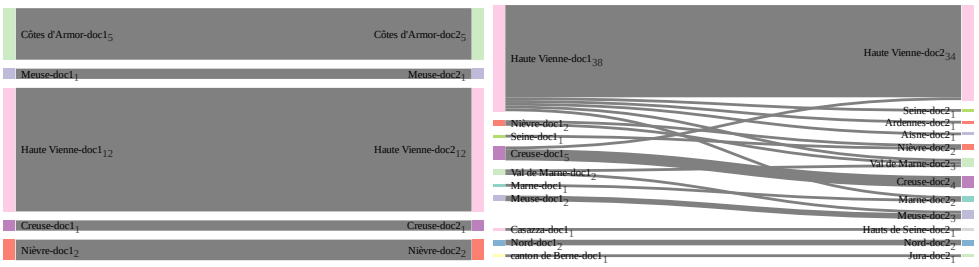


Figure 4.13 – Sankey diagrams showing the migration of people in the 18th and 19th centuries, extracted from their birth and death places.

4.5.3 . Sociology thesis in France

We describe in this third use case how ComBiNet can be used to answer questions about thesis in France between 2016 and 2022. Indeed, some sociological datasets made of documents can also be well modeled as bipartite multivariate dynamic networks like for example thesis dissertations: a thesis is a document with specific attributes such as the subject, the doctoral school, the domain, the university, and the date of defense, and mention several peoples who are socially connected through the thesis defense with different roles: author (*auteur* in french), director(s) (*directeur*), referees (*rapporteur*), and jury president (*président de jury*). We present here an exploration of the data by ourselves using ComBiNet. A first look at the graph measures tells us that 896 theses have been defended in sociology in France between 2016 and 2021 in France, with 2453 persons included in the defenses (see Figure 4.14 bottom). The bipartite node-link view shows us an overview of the network but is hard to parse due to the network's size. Zoom actions though allow centering the view for specific parts of the network. The map view allows us to see that thesis has been defended all around France. We can however see that the majority of theses are defended in Paris. This is confirmed

if we look at the distribution of the cities (Figure 4.14 bottom right): around half of the defenses are in Paris, compared to the rest of the country which is more or less homogeneous. By setting the threshold to link creation to one (meaning that a link is created between two documents if they mention at least one common person), a lot of links are created as seen in Figure 4.14 (right). It means that a lot of thesis defenses include referees and juries from different cities.

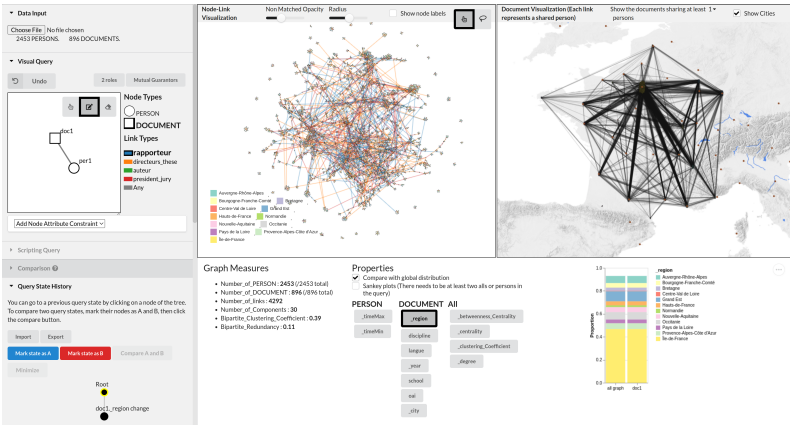


Figure 4.14 – ComBiNet used for exploring theses of sociology defended in France between 2016 and 2021. The bipartite and map views show an overview of two visions of the network. The user selects the *region* attribute, showing the geographical distribution of the defended thesis.

Let's now try to answer an interesting question: "Do referees and jury presidents often ask thesis directors to be referees and jury presidents in their turn of another thesis where they are directors ?". For this, we can construct a visual query representing this pattern by creating two person nodes and two document nodes, and by connecting them with two president links and two referee or jury director links in a symmetrical way, as shown in Figure 4.15 (right). The occurrence table tells us that this pattern has been found 76 times in the network, meaning that this is a recurrent behavior. We are now interested in characterizing the thesis occurring in this pattern, by their regions. We can look at the *city* attribute distribution for this thesis by selecting it in the attribute view as shown in Figure 4.15 (bottom right). We can first see on the map that this pattern occurs mainly in the biggest cities of the country. By selecting the Sankey view option, we can investigate if this pattern occurs between thesis defended in different regions or if it occurs mainly in the same ones. We learn that it depends mainly on the regions: in Bourgogne-Fanche-Comté 26 out of 29 theses are connected with the thesis of another region. On contrary, in *Occitanie* it is the case for only 4 out of 17. On average, we can see that this pattern occurs a lot for theses of the same region. In Ile-de-France, it is the case for around half of the thesis (28/50). This exploratory

analysis shows that ComBiNet can be used to explore and gain insight into such datasets.

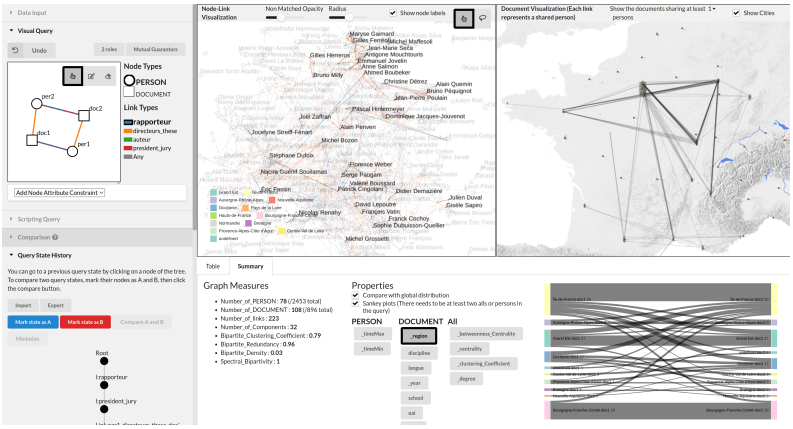


Figure 4.15 – Sociology thesis dataset explored with ComBiNet. The user constructed a visual query to see if there are symmetrical relationships between thesis directors and referees (or jury directors). The *region* attribute is selected with the Sankey option, letting the user see if there are correlations between the regions of the thesis found in this pattern.

4.6 . Formative Usability Study

We performed a formative usability study with two historians and one expert in visualization. We had 3 meetings with each and gave them control of the tool to see if they could use it to explore their data and perform queries and comparisons. At each meeting, we asked them to speak aloud, commenting on their aims and actions. At the end of each session, we asked them their general feedback and what other features they would like to have. We improved the system and made the changes asked by the users before setting up new appointments. This usability study led to the redesign of some core features, like the activation of the comparison mode which is now started by first marking the state nodes in the provenance tree. It also led to the implementation of new features, such as the person and document tables (which are updated after each query), the persistent selection of nodes across the two views and the tables, and the undo feature for visual queries. At the final meetings, the three users were able to perform exploration, queries, and comparisons to answer socio-historical questions by themselves.

4.6.1 . Feedback

All three users liked the table views and were exploiting them to study in depth who were the person and documents found in their specific queries. Both historians

liked the Sankey diagram of the attributes, allowing them to see the evolution of distributions and answering several of their questions. Our collaborator of the use case #2 was making sense of it by linking the migration patterns she was seeing in the Sankey diagram with specific persons of the dataset she knew in depth. She was also curious about other migration patterns she was not aware of and wanted to know who these persons were, the system allowing her to select them and follow a deeper exploration.

4.7 . Discussion

Query Expressiveness. The visual query system currently allows finding occurrences of attributed subgraphs, with potential union operations on constraints (links and node attribute values can be set at one value or as a set of values). Being able to express attribute constraints (other than for labels and ids) and unions is new compared to other visual graph query systems. More complex constraints are then expressible using the Cypher editor, such as dependent constraints, e.g., if one node attribute value has to be greater or lower than another attribute value. The visual query system could be extended by introducing more complex time constraints capabilities, such as in [93].

Scalability. We assess the scalability in network size (number of nodes and links) concerning the cluttering and readability of the network visualizations. Our biggest dataset from #3 comprises 7212 nodes (4886 persons and 2326 events) and 7790 links, after splitting the documents into birth and marriage event nodes. The system allows the exploration of networks of this size with a decent frame rate. ComBiNet allows navigating relatively large sparse graphs (thousands of nodes) with the node-link visualization using zoom & pan and filtering with the query system. It lets users focus on subsets of the data, one or two at a time.

Generalizibility. The system has been designed specifically for bipartite multivariate dynamic networks, which models well a diversity of historical sources we encountered via our collaborations: marriage acts, birth/death certificates, construction/work contracts, census, and migrations forms. Moreover, bipartite multivariate dynamic network can also be used to model other similar data types, such as scientific publications or thesis data. However, other kinds of historical textual data exist where documents can mention each other, such as in private letters for example. The model and interface would need to be slightly modified to take into account document-to-document links for these datasets. Bipartite networks are also used in various other disciplines, such as biology [73] and chemistry [74]. ComBiNet could be extended to these other application domains, in particular by modifying the map view to show other location data related to the entities of the network, or removing it altogether if it makes no sense for a particular domain.

4.8 . Conclusion and Future Work

We presented ComBiNet, a system for exploring social networks modeled from historical textual sources, aimed at social scientists. It relies on modeling data as bipartite, multivariate, dynamic social networks where persons are linked to documents or events using typed links that express roles. Our tool ComBiNet relies on this data model to let historians explore their data and then answer their socio-historical questions using 1) dynamic queries on the network structure and attributes to highlight groups of interests, and 2) visual comparisons to contrast selected groups according to their structure, time, or any other attribute. The results can be visualized as a node-link diagram, a geographical map, graph measures, and distributions of values for the attributes. We have shown that complex explorations and analyses were easy or possible to perform, and validated our approach by first describing two use cases among many more projects we are collaborating with and by performing a formative usability study showing that the system is usable by social scientists.

By specifying a unifying data model and novel high-level visual and interactive tools for comparing topology, attributes, and time, social scientists were able to clean their data more easily by finding errors and inconsistencies by exploring the network and querying errors-induced patterns. Thanks to the document-centered model, it was easy for them to trace back the errors and inconsistencies to the sources for corrections. With the same representation, they were able to operate explorations and analyses using complex interactions implemented in ComBiNet such as coordinated views, visual querying, and comparison mechanisms.

Using these mechanisms, social scientists were able to perform visual exploratory analyses of their network based on topological and attribute descriptions and comparisons of subgroups of interests, and of the overall network. This methodology allows them to either ground or refute their hypotheses in their results, or to generate new ones from new insight revealed thanks to the complex exploratory and interaction mechanisms.

We believe ComBiNet leads the way toward a new generation of highly interactive exploration tools applicable to wrangle and analyze a wide variety of real social networks modeled from textual sources, with a focus on the traceability of the network and results, which is essential for any historical workflow.

For future work, ComBiNet could be extended to support more SNA measures and computations such as clustering; it would create a new attribute containing a cluster identifier. The interface currently proposes two layouts based on the topology and the geolocations of the entities. Providing more layout options could be interesting, especially one to highlight better the time, similar to the PAOHvis technique [142]. Finally, the interface currently lets social scientists build their queries to answer questions they have about their data. In the future, the system could make suggestions on the query construction process with a mixed-initiative perspective, to guide users towards frequent subgraphs in the data which could be interesting to investigate.

Bibliography

[1] NodeXL: Simple network analysis for social media.

[2] Mashael Alkadi, Vanessa Serrano, James Scott-Brown, Catherine Plaisant, Jean-Daniel Fekete, Uta Hinrichs, and Benjamin Bach. Understanding barriers to network exploration with visualization: A report from the trenches. *IEEE Transactions on Visualization and Computer Graphics*, 27(2), 2022.

[3] Keith Andrews, Martin Wohlfahrt, and Gerhard Wurzinger. Visual Graph Comparison. In *2009 13th International Conference Information Visualisation*, pages 62–67, July 2009.

[4] F. J. Anscombe. Graphs in Statistical Analysis. *The American Statistician*, 27(1):17–21, February 1973.

[5] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. In Eytan Adar, Matthew Hurst, Tim Finin, Natalie S. Glance, Nicolas Nicolov, and Belle L. Tseng, editors, *ICWSM*. The AAAI Press, 2009.

[6] Sugato Basu, Ian Davidson, and Kiri Wagstaff. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC, first edition, 2008.

[7] Leilani Battle and Jeffrey Heer. Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau. *Computer Graphics Forum*, 38(3):145–159, 2019.

[8] Jacques Bertin. *Sémiologie graphique: les diagrammes, les réseaux, les cartes*. Paris: Gauthier-Villars, 1967.

[9] A. Bezerianos, F. Chevalier, P. Dragicevic, N. Elmqvist, and J.d. Fekete. GraphDice: A System for Exploring Multivariate Social Networks. *Computer Graphics Forum*, 29(3):863–872, 2010.

[10] Marc Bloch. *Apologie Pour l'histoire*. A. Colin, 1949.

[11] Christian Böhm and Claudia Plant. HISSCLU: A hierarchical density-based method for semi-supervised clustering. In *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '08, pages 440–451, New York, NY, USA, 2008. ACM.

[12] Christian Bors, John Wenskovitch, Michelle Dowling, Simon Attfield, Leilani Battle, Alex Endert, Olga Kulyk, and Robert S. Laramee. A Provenance

Task Abstraction Framework. *IEEE Computer Graphics and Applications*, 39(6):46–60, November 2019.

[13] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, December 2011.

[14] Pierre Bourdieu. Sur les rapports entre la sociologie et l'histoire en Allemagne et en France. *Actes de la Recherche en Sciences Sociales*, 106(1):108–122, 1995.

[15] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On Modularity Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188, February 2008.

[16] Peter Burke. *History and Social Theory*. Polity, 2005.

[17] Steven P. Callahan, Juliana Freire, Emanuele Santos, Carlos E. Scheidegger, Cláudio T. Silva, and Huy T. Vo. VisTrails: Visualization meets data management. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data - SIGMOD '06*, page 745, Chicago, IL, USA, 2006. ACM Press.

[18] Charles-Olivier Carbonell. *L'Historiographie*. FeniXX, January 1981.

[19] Stuart-K. Card, Jock-D. Mackinlay, and Ben Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers In, San Francisco, Calif, February 1999.

[20] Duen Horng Chau, Christos Faloutsos, Hanghang Tong, Jason I. Hong, Brian Gallagher, and Tina Eliassi-Rad. GRAPHITE: A Visual Query System for Large Graphs. In *2008 IEEE International Conference on Data Mining Workshops*, pages 963–966, December 2008.

[21] J. S. Coleman. Introduction to mathematical sociology. *Introduction to mathematical sociology.*, 1964.

[22] TEI Consortium. TEI P5: Guidelines for electronic text encoding and interchange, February 2021.

[23] Pascal Cristofoli. Aux sources des grands réseaux d'interactions. *Reseaux*, 152(6):21–58, 2008.

[24] Pascal Cristofoli. Principes et usages des dessins de réseaux en SHS. *La visualisation des données en histoire*, page 35, 2015.

[25] Pascal Cristofoli and Nicoletta Rolla. Temporalités à l'œuvre dans les chantiers du bâtiment. *Temporalités. Revue de sciences sociales et humaines*, (27), June 2018.

[26] Tarik Crnovrsanin, Chris W. Muelder, Robert Faris, Diane Felmlee, and Kwan-Liu Ma. Visualization techniques for categorical analysis of social networks with multiple edge sets. *Social Networks*, 37:56–64, 2014.

[27] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[28] Erick Cuenca, Arnaud Sallaberry, Dino Ienco, and Pascal Poncelet. VERTIGo: A Visual Platform for Querying and Exploring Large Multilayer Networks. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2021.

[29] Zach Cutler, Kiran Gadhave, and Alexander Lex. Trrack: A Library for Provenance-Tracking in Web-Based Visualizations. In *2020 IEEE Visualization Conference (VIS)*, pages 116–120, October 2020.

[30] Mandeep K. Dhami, Ian K. Belton, and David R. Mandel. The "analysis of competing hypotheses" in intelligence analysis. *Applied Cognitive Psychology*, 33(6):1080–1090, 2019.

[31] Jana Diesner, Craig Evans, and Jinseok Kim. Impact of Entity Disambiguation Errors on Social Network Properties. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):81–90, 2015.

[32] Dana Diminescu. The migration of ethnic germans from romania to west germany: Insights from the archives of the former communist regime. In *CERS, Public Lecture, UCLA*, Los Angeles, United States, March 2020.

[33] Nicole Dufournaud. La recherche empirique en histoire à l'ère numérique. *Gazette des archives*, 240(4):397–407, 2015.

[34] Nicole Dufournaud. Comment rendre visible le rôle économique des femmes sous l'Ancien Régime ? Étude méthodologique sur les marchandes à Nantes aux XVIe et XVIIe siècles. In Bernard Michon and Nicole Dufournaud, editors, *Femmes et Négoce Dans Les Ports Européens (Fin Du Moyen Age - XIXe Siècle)*, pages 65–84. Peter Lang, 2018.

[35] Nicole Dufournaud and Jean-Daniel Fekete. Comparaison d'outils pour la visualisation de sources historiques codées en XML/TEI. *Document numérique*, 9(2):37–56, April 2006.

[36] Cody Dunne, Nathalie Henry Riche, Bongshin Lee, Ronald Metoyer, and George Robertson. GraphTrail: Analyzing large multivariate, heterogeneous

networks while supporting exploration history. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1663–1672, New York, NY, USA, May 2012. Association for Computing Machinery.

[37] P. Erdös and A. Rényi. On the evolution of random graphs. In *On the Evolution of Random Graphs*, pages 38–82. Princeton University Press, October 2011.

[38] Emily Erikson and Peter Bearman. Malfeasance and the Foundations for Global Trade: The Structure of English Trade in the East Indies, 1601–1833. *American Journal of Sociology*, 112(1):195–230, July 2006.

[39] Michael Eve. Deux traditions d'analyse des reseaux sociaux. *Réseaux*, 115(5):183–212, 2002.

[40] Lucien Febvre. VERS UNE AUTRE HISTOIRE. *Revue de Métaphysique et de Morale*, 54(3/4):225–247, 1949.

[41] Jean-Daniel Fekete, Danyel Fisher, Arnab Nandi, and Michael Sedlmair. *Progressive Data Analysis and Visualization*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, April 2019.

[42] L. Freeman. Visualizing Social Networks. *J. Soc. Struct.*, 2000.

[43] L.C. Freeman. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press, 2004.

[44] Manuel Freire, Catherine Plaisant, Ben Shneiderman, and Jen Golbeck. ManyNets: An interface for multiple network analysis and visualization. In *CHI '10*, CHI '10, pages 213–222, New York, NY, USA, 2010. ACM.

[45] Michael Friendly. Visions and Re-Visions of Charles Joseph Minard. *Journal of Educational and Behavioral Statistics*, 27(1):31–51, March 2002.

[46] Michael Friendly. A Brief History of Data Visualization. In Chun-houh Chen, Wolfgang Härdle, and Antony Unwin, editors, *Handbook of Data Visualization*, Springer Handbooks Comp.Statistics, pages 15–56. Springer, Berlin, Heidelberg, 2008.

[47] GEDCOM: The genealogy data standard.

[48] Mohammad Ghoniem, J.-D. Fekete, and Philippe Castagliola. A comparison of the readability of graphs using node-link and matrix-based representations. In *IEEE Symposium on Information Visualization*, pages 17–24. Ieee, 2004.

[49] Carlo Ginzburg and Carlo Poni. La micro-histoire. *Le Débat*, 17(10):133, 1981.

[50] Barney G. Glaser and Anselm L. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Transaction, New Brunswick, 5. paperback print edition, 2010.

[51] Michael Gleicher. Considerations for visualizing comparison. *IEEE Trans. Vis. Comput. Graphics*, 24(1):413–423, 2018.

[52] Claudia Goldin. Cliometrics and the Nobel. *Journal of Economic Perspectives*, 9(2):191–208, June 1995.

[53] Martin Grandjean. Social network analysis and visualization: Moreno's Sociograms revisited, 2015.

[54] Martin Grandjean. Analisi e visualizzazioni delle reti in storia. L'esempio della cooperazione intellettuale della Società delle Nazioni. *ME*, (2/2017), 2017.

[55] Maurizio Gribaudi and Alain Blum. Des catégories aux liens individuels : l'analyse statistique de l'espace social. *Annales*, 45(6):1365–1402, 1990.

[56] Jo Guldi and David Armitage. *The History Manifesto*. Cambridge University Press, October 2014.

[57] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11–15, Pasadena, CA USA, 2008.

[58] Klaus Hamberger, Cyril Grange, Michael Houseman, and Christian Momon. Scanning for patterns of relationship: Analyzing kinship and marriage networks with Puck 2.0. *The History of the Family*, 19(4):564–596, October 2014.

[59] Klaus Hamberger, Michael Houseman, and R. White, Douglas. Kinship network analysis. In John Scott & Peter J. Carrington, editor, *The Sage Handbook of Social Network Analysis*, pages 533–549. Sage Publications, 2011.

[60] Mountaz Hascoët and Pierre Dragicevic. Interactive graph matching and visual comparison of graphs and clustered graphs. In Genny Tortora, Stefano Levialdi, and Maurizio Tucci, editors, *AVI '12*, pages 522–529. ACM, 2012.

[61] J. Heer and D. Boyd. Vizster: Visualizing online social networks. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 32–39, October 2005.

[62] Jeffrey Heer. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, 116(6):1844–1850, 2019.

[63] Louis Henry and Michel Fleury. Des registres paroissiaux a l'histoire de la population: Manuel de dépouillement et d'exploitation de l'état civil ancien. *Population (French Edition)*, 11(1):142–144, 1956.

[64] Nathalie Henry, Jean-Daniel Fekete, and Michael J. McGuffin. NodeTrix: A Hybrid Visualization of Social Networks. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1302–1309, November 2007.

[65] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, and Sabrina Kirrane et al. Knowledge graphs. *ACM Comput. Surv.*, 54(4), July 2021.

[66] Pat Hudson and Mina Ishizu. *History by Numbers: An Introduction to Quantitative Approaches*. Bloomsbury Publishing, November 2016.

[67] Infovis SC policies FAQ.

[68] Piers J Ingram, Michael PH Stumpf, and Jaroslav Stark. Network motifs: Structure does not determine function. *BMC Genomics*, 7:108, May 2006.

[69] Frédéric Kaplan. The Venice Time Machine. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, DocEng '15, page 73, New York, NY, USA, September 2015. Association for Computing Machinery.

[70] Karine Karila-Cohen, Claire Lemercier, Isabelle Rosé, and Claire Zalc. Nouvelles cuisines de l'histoire quantitative. *Annales. Histoire, Sciences Sociales*, 73(4):773–783, December 2018.

[71] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual Analytics: Definition, Process, and Challenges. In Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North, editors, *Information Visualization: Human-Centered Issues and Perspectives*, Lecture Notes in Computer Science, pages 154–175. Springer, Berlin, Heidelberg, 2008.

[72] Florian Kerschbaumer, Linda von Keyserlingk-Rehbein, Martin Stark, and Marten Düring. *The Power of Networks. Prospects of Historical Network Research*. Routledge, December 2021.

[73] Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and cellular networks. *PLoS computational biology*, 5(5):e1000385, 2009.

[74] Elena V. Konstantinova and Vladimir A. Skorobogatov. Application of hypergraph theory in chemistry. *Discrete Mathematics*, 235(1-3):365–383, May 2001.

[75] C. Kosak, J. Marks, and S. Shieber. Automating the layout of network diagrams with specified visual organization. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(3):440–454, March 1994.

[76] Ernest Labrousse. *La Crise de l'économie Française à La Fin de l'Ancien Régime et Au Début de La Révolution*, volume 1. Presses Universitaires de France-PUF, 1990.

[77] Charles-Victor Langlois and Charles Seignobos. *Introduction aux études historiques*. ENS Éditions, February 2014.

[78] Katherine A. Larson. Thomas F. Tartaron, Maritime Networks in the Mycenaean World. New York: Cambridge University Press, 2013. *Comparative Studies in Society and History*, 56(4):1064–1065, October 2014.

[79] Emmanuel Lazega. *Réseaux sociaux et structures relationnelles*. Presses universitaires de France, Paris, 1998.

[80] Claire Lemercier. 12. Formal network methods in history: Why and how? In Georg Fertig, editor, *Social Networks, Political Institutions, and Rural Societies*, volume 11, pages 281–310. Brepols Publishers, Turnhout, January 2015.

[81] Claire Lemercier and Claire Zalc. *Quantitative Methods in the Humanities: An Introduction*. University of Virginia Press, March 2019.

[82] Claire Lemercier and Claire Zalc. Back to the Sources: Practicing and Teaching Quantitative History in the 2020s. *Capitalism*, 2(2):473–508, 2021.

[83] Bernard Lepetit. L'histoire quantitative : deux ou trois choses que je sais d'elle. *Histoire & Mesure*, 4(3):191–199, 1989.

[84] Carola Lipp. Kinship Networks, Local Government, and Elections in a Town in Southwest Germany, 1800-1850. *Journal of Family History*, 30(4):347–365, October 2005.

[85] Gribaudi Maurizio. *Espaces, Temporalités, Stratifications :. Exercices Méthodologiques Sur Les Réseaux Sociaux*. Editions de l'Ecole des Hautes Etudes en Sciences Sociales, Paris, January 2000.

[86] Philip Mayer. Migrancy and the Study of Africans in Towns. *American Anthropologist*, 64(3):576–592, 1962.

[87] Fintan McGee, Benjamin Renoust, Daniel Archambault, Mohammad Ghoniem, Andreas Kerren, and Bruno Pinaud et al. *Visual Analysis of Multilayer Networks*. Synthesis Lectures on Visualization. Morgan & Claypool Publishers, 2021.

[88] Michael J. McGuffin. Simple algorithms for network visualization: A tutorial. *Tsinghua Science and Technology*, 17(4):383–398, August 2012.

[89] Pierre Mercklé and Claire Zalc. Peut-on modéliser la persécution ?: Apports et limites des approches quantifiées sur le terrain de la Shoah. *Annales. Histoire, Sciences Sociales*, 73(4):923–957, December 2018.

[90] R. Michalski, P. Kazienko, and D. Krol. Predicting Social Network Measures Using Machine Learning Approach. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1056–1059, Istanbul, August 2012. IEEE.

[91] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594):824–827, October 2002.

[92] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. Lulu.com, 2019.

[93] Megan Monroe, Krist Wongsuphasawat, Catherine Plaisant, Ben Shneiderman, Jeff Millstein, and Sigfried Gold. Exploring point and interval event patterns: Display methods and interactive visual query. *University of Maryland Technical Report*, 2012.

[94] J. L. Moreno. *Who Shall Survive?: A New Approach to the Problem of Human Interrelations*. Who Shall Survive?: A New Approach to the Problem of Human Interrelations. Nervous and Mental Disease Publishing Co, Washington, DC, US, 1934.

[95] J. L. Moreno. Foundations of Sociometry: An Introduction. *Sociometry*, 4(1):15, February 1941.

[96] Zacarias Moutoukias. Buenos Aires, port between two oceans: Mobilities, networks, stratifications (2nd half of the 18th century). *E-SPANIA-REVUE ELECTRONIQUE D ETUDES HISPANIQUES MEDIEVALES*, 25, 2016.

[97] Zacharias Moutoukias. Réseaux personnels et autorité coloniale : Les négociants de Buenos Aires au XVIIIe siècle. *Annales. Histoire, Sciences Sociales*, 47(4-5):889–915, October 1992.

[98] Andrej Mrvar and Vladimir Batagelj. Analysis and visualization of large networks with program package Pajek. *Complex Adaptive Systems Modeling*, 4(1), April 2016.

[99] Natural earth.

[100] Neo4j graph data platform.

[101] Rolla Nicoletta. Mobilité et conflits. Travailler sur les chantiers de construction piémontais dans la première moitié du XVIIIe siècle. In Andrea Caracausi and Marco Schnyder, editors, *Travail et Mobilité En Europe (XVIe-XIXe Siècles)*, Coll. Histoire et Civilisations. Presses universitaires du Septentrion, Villeneuve d'Ascq, 2018.

[102] Carolina Nobre, Marc Streit, and Alexander Lex. Juniper: A Tree+Table Approach to Multivariate Graph Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):544–554, January 2019.

[103] Gérard Noiriel. Naissance du métier d'historien. *Genèses. Sciences sociales et histoire*, 1(1):58–85, 1990.

[104] Maryjane Osa. *Solidarity And Contention: Networks Of Polish Opposition*. Univ Of Minnesota Press, Minneapolis, first edition edition, July 2003.

[105] John F. Padgett and Christopher K. Ansell. Robust Action and the Rise of the Medici, 1400-1434. *American Journal of Sociology*, 98(6):1259–1319, May 1993.

[106] Pajek — Analysis and visualization of very large networks.

[107] Terence J. Parr and Russell W. Quong. ANTLR: A predicated-LL (k) parser generator. *Software: Practice and Experience*, 25(7):789–810, 1995.

[108] Vanessa Peña-Araya, Tong Xue, Emmanuel Pietriga, Laurent Amsaleg, and Anastasia Bezerianos. HyperStorylines: Interactively untangling dynamic hypergraphs. *Information Visualization*, 21(1):38–62, January 2022.

[109] Cindarella Sarah Maria Petz. *On Combining Network Research and Computational Methods on Historical Research Questions and Its Implications for the Digital Humanities*. PhD thesis, Technische Universität München, 2022.

[110] James P. Philips and Nasseh Tabrizi. Historical Document Processing: Historical Document Processing: A Survey of Techniques, Tools, and Trends, September 2020.

[111] Robert Pienta, Fred Hohman, Alex Endert, Acar Tamersoy, Kevin Roundy, Chris Gates, Shamkant Navathe, and Duen Horng Chau. VIGOR: Interactive Visual Exploration of Graph Query Results. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):215–225, January 2018.

[112] Alexis Pister, Nicole Dufournaud, Pascal Cristofoli, Christophe Prieur, and Jean-Daniel Fekete. From Historical Documents To Social Network Visualization: Potential Pitfalls and Network Modeling. In *7th Workshop on Visualization for the Digital Humanities (VIS4DH)*, 2022.

[113] Antoine Prost. *Douze Leçons sur l'histoire*. Média Diffusion, April 2014.

[114] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[115] Eric Ragan, Endert Alex, Jibonananda Sanyal, and Jian Chen. Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes. *IEEE Transactions on Visualization and Computer Graphics*, 22(1), January 2016.

[116] Ramana Rao and Stuart K. Card. The table lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '94, pages 318–322, New York, NY, USA, 1994. Association for Computing Machinery.

[117] Donghao Ren, Bongshin Lee, and Matthew Brehmer. Charticulator: Interactive Construction of Bespoke Chart Layouts. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):789–799, January 2019.

[118] Pedro Ribeiro and Fernando Silva. Discovering Colored Network Motifs. In Pierluigi Contucci, Ronaldo Menezes, Andrea Omicini, and Julia Poncela-Casasnovas, editors, *Complex Networks V*, Studies in Computational Intelligence, pages 107–118, Cham, 2014. Springer International Publishing.

[119] Giulio Rossetti and Rémy Cazabet. Community discovery in dynamic networks: A survey. *ACM Comput. Surv.*, 51(2), February 2018.

[120] Fabrice Rossi, Nathalie Vialaneix, and Florent Hautefeuille. Exploration of a large database of French notarial acts with social network methods. *Digital Medievalist*, 9:2013, July 2014.

[121] Juan A. Rubio-Mondejar and Josean Garrues-Irurzun. Women entrepreneurs and family networks in Andalusia (Spain) during the second industrial revolution. *Business History*, pages 1–22, May 2022.

[122] C.J. Rueda and Catedral de Buenos Aires. *Matrimonios de La Catedral de Buenos Aires, 1747-1823*. Number v. 2 in Fuentes Históricas y Genealógicas Argentinas. Fuentes Históricas y Genealógicas Argentinas, 1989.

[123] Anni Sairio. Methodological and practical aspects of historical network analysis: A case study of the Bluestocking letters. In Arja Nurmi, Minna Nevala,

and Minna Palander-Collin, editors, *Pragmatics & Beyond New Series*, volume 183, pages 107–135. John Benjamins Publishing Company, Amsterdam, 2009.

[124] Bahador Saket, Paolo Simonetto, and Stephen Kobourov. Group-level graph visualization taxonomy. In N. Elmqvist, M. Hlawitschka, and J. Kennedy, editors, *EuroVis - Short Papers*. The Eurographics Association, 2014.

[125] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. Vega-lite: A grammar of interactive graphics. *IEEE Trans. Vis. Comput. Graphics*, 23(1):341–350, 2016.

[126] Shrutika S. Sawant and Manoharan Prabukumar. A review on graph-based semi-supervised learning methods for hyperspectral image classification. *The Egyptian Journal of Remote Sensing and Space Science*, 2018.

[127] John Scott. Social Network Analysis. *Sociology*, 22(1):109–127, February 1988.

[128] Vanessa Serrano Molinero, Benjamin Bach, Catherine Plaisant, Nicole Dufournaud, and Jean-Daniel Fekete. Understanding the use of the vistorian: Complementing logs with context mini-questionnaires. In *Visualization for the Digital Humanities Workshop*, Phoenix, United States, October 2017.

[129] Rachel Shadoan and Chris Weaver. Visual Analysis of Higher-Order Conjunctive Relationships in Multidimensional Data Using a Hypergraph Query System. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2070–2079, December 2013.

[130] Termeh Shafie, David Schoch, Jimmy Mans, Corinne Hofman, and Ulrik Brandes. Hypergraph Representations: A Study of Carib Attacks on Colonial Forces, 1509-1700. *Journal of Historical Network Research*, pages 52–70 Pages, October 2017.

[131] Ben Shneiderman. Dynamic queries for visual information seeking. *IEEE Softw.*, 11(6):70–77, November 1994.

[132] Georg Simmel. *Soziologie: Untersuchungen über die Formen der Vergesellschaftung*. Duncker & Humblot, Berlin, 7. aufl edition, 2013.

[133] Marc A. Smith, Ben Shneiderman, Natasa Milic-Frayling, Eduarda Mendes Rodrigues, Vladimir Barash, Cody Dunne, Tony Capone, Adam Perer, and Eric Gleave. Analyzing (social media) networks with NodeXL. In John M. Carroll, editor, *Proceedings of the Fourth International Conference on Communities and Technologies, C&T 2009, University Park, PA, USA, June 25-27, 2009*, pages 255–264. ACM, 2009.

[134] SNA — Tools for social network analysis.

[135] John Snow. On the Mode of Communication of Cholera. *Edinb Med J*, 1(7):668–670, January 1856.

[136] John T. Stasko, Carsten Görg, and Zhicheng Liu. Jigsaw: Supporting investigative analysis through interactive visualization. *Inf. Vis.*, 7(2):118–132, 2008.

[137] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

[138] Shazia Tabassum, Fabiola S. F. Pereira, Sofia Fernandes, and João Gama. Social network analysis: An overview. *WIREs Data Mining and Knowledge Discovery*, 8(5):e1256, 2018.

[139] Natkamon Tovanich, Alexis Pister, Gaelle Richer, Paola Valdivia, Christophe Prieur, Jean-Daniel Fekete, and Petra Isenberg. VAST 2020 Contest Challenge: GraphMatchMaker: Visual Analytics for Graph Comparison and Matching. *IEEE Computer Graphics and Applications*, pages 1–1, 2021.

[140] John W. Tukey. The Future of Data Analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962.

[141] John W. Tukey. *Exploratory Data Analysis*. Pearson, Reading, Mass, 1er édition edition, January 1977.

[142] Paola Valdivia, Paolo Buono, Catherine Plaisant, Nicole Dufournaud, and Jean-Daniel Fekete. Analyzing Dynamic Hypergraphs with Parallel Aggregated Ordered Hypergraph Visualization. *IEEE Trans. Visual. Comput. Graphics*, 27(1):1–13, January 2021.

[143] Guido van Rossum. Python tutorial. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995.

[144] Ingeborg van Vugt. Using multi-layered networks to disclose books in the republic of letters. *Journal of Historical Network Research*, 1(1):25–51, October 2017.

[145] Corinna Vehlow, Fabian Beck, and Daniel Weiskopf. The state of the art in visualizing group structures in graphs. In R. Borgo, F. Ganovelli, and I. Viola, editors, *Eurographics Conference on Visualization (EuroVis) - STARs*. The Eurographics Association, 2015.

[146] VisMaster: Visual analytics — Mastering the information age.

[147] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al. Constrained k-means clustering with background knowledge. In *Icml*, volume 1, pages 577–584, 2001.

[148] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, November 1994.

[149] Charles Wetherell. Historical Social Network Analysis. *Int Rev of Soc His*, 43(S6):125–144, December 1998.

[150] Kai Xu, Alvitta Ottley, Conny Walchshofer, Marc Streit, Remco Chang, and John Wenskovitch. Survey on the Analysis of User Interactions and Visualization Provenance. *Computer Graphics Forum*, 39(3):757–783, June 2020.

[151] Michelle X. Zhou. "Big picture": Mixed-initiative visual analytics of big data. In *Proceedings of the 6th International Symposium on Visual Information Communication and Interaction*, VINCI '13, page 120, New York, NY, USA, 2013. Association for Computing Machinery.