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 Social Networks: Traceability, Exploration, and Analysis

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

1 Introduction **7**
1.1 Social History and Historical Social Network Analysis 8
1.2 Visualization and Visual Analytics 9
1.3 Historical Social Networks Visual Analytics 12
1.4 Contributions and Research Statement 12

2 Related Work **15**
2.1 Visualization 15
2.1.1 Information Visualization 16
2.1.2 Visual Analytics 18
2.2 Quantitative Social History 20
2.2.1 History, Social History, and Methodology 20
2.2.2 Quantitative History 21
2.2.3 Digital Humanities 22
2.3 Historical Social Network Analysis 25
2.3.1 Sociometry to SNA 25
2.3.2 Methods and Measures 26
2.3.3 Historical Social Network Analysis 29
2.3.4 Network Modeling 29
2.4 Social Network Visualization 32
2.4.1 Graph Drawing 32
2.4.2 Social Network Visual Analytics 33

3 HSNA Process and Network Modeling **33**
3.1 Context 33
3.2 Related Work 34
3.2.1 History Methodology 35
3.2.2 Historian Workflows 35
3.3 Historical Social Network Analysis Workflow 36
3.3.1 Textual Sources Acquisition 37
3.3.2 Digitization 37
3.3.3 Annotation 38
3.3.4 Network Creation 38
3.3.5 Network Analysis and Visualization 39
3.4 Network modeling and analysis 39
3.4.1 Network Models 40
3.4.2 Bipartite Multivariate Dynamic Social Network 42
3.4.3 Examples 43

3.5 Applications 44
3.6 Discussion 46
3.7 Conclusion 46

4 ComBiNet: Visual Query and Comparison of Bipartite Dynamic Multivariate Networks with Roles 49
4.1 Context 49
4.2 Related Work 51
4.2.1 Graphlet Analysis 51
4.2.2 Visual Graph Querying 52
4.2.3 Visual Graph Comparison 53
4.2.4 Provenance 53
4.3 Task Analysis and Design Process 53
4.3.1 Use Cases 54
4.3.2 Tasks Analysis 57
4.4 The ComBiNet System 58
4.4.1 Visualizations 59
4.4.2 Query Panel 60
4.4.3 Comparison 66
4.4.4 Implementation 69
4.5 Use Cases 69
4.5.1 Construction sites in Piedmont (#1) 69
4.5.2 French Genealogy (#2) 70
4.5.3 Sociology thesis in France 72
4.6 Formative Usability Study 74
4.6.1 Feedback 74
4.7 Discussion 75
4.8 Conclusion and Future Work 75

5 PK-Clustering 77
5.1 Context 77
5.2 Related Work 80
5.2.1 Graph Clustering 81
5.2.2 Semi-supervised Clustering 81
5.2.3 Mixed-Initiative Systems and Interactive Clustering 82
5.2.4 Groups in Network Visualization 82
5.2.5 Ensemble Clustering 82
5.2.6 Summary 83
5.3 PK-clustering 83
5.3.1 Overview 84
5.3.2 Specification of Prior Knowledge 85
5.3.3 Running the Clustering Algorithms 85
5.3.4 Matching Clustering Results and Prior Knowledge 86

5.3.5 Ranking the Algorithms . . . 87
5.3.6 Reviewing the Ranked List of Algorithms . . . 88
5.3.7 Reviewing and Consolidating Final Results . . . 89
5.3.8 Wrapping up and Reporting Results . . . 93
5.4 Case studies . . . 94
5.4.1 Marie Boucher Social Network . . . 95
5.4.2 Lineages at VAST . . . 95
5.4.3 Feedback from practitioners . . . 98
5.5 Discussion . . . 99
5.5.1 Limitations . . . 100
5.5.2 Performance . . . 100
5.6 Conclusion . . . 101

6 Conclusion **103**
6.1 Summary . . . 103
6.2 Discussion . . . 103
6.3 Perspectives . . . 105
6.4 Conclusion . . . 107

2 Related Work

Social historians rely on textual historical documents to study social groups through their structures and socio-economic place in societies of the past [?]. They read and analyze documents they can find from a period and subject of interest, and make their conclusions through deep inspection and cross-referencing of 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 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 [?, ?]. This approach is directly inspired by SNA, which is a well-known method that sociologist theorized to understand and describe real world social relationships modeled as networks [?, ?]. Historians appropriated this method, by extracting relationships from historical documents. The specificy of HSNA in contrast of its sociology counterpart is therefore the modeling of the network from the historical documents—which are at the core of the historical work [?]—and the integration of the time aspect which is often disregarded in traditional SNA. Once they successfully constructed a network—which is a long and tedious process—they typically use network measures and visualization techniques to confirm or generate new hypotheses [?]. Visualization let them unfold the structure of their data, revealing potentially interesting social patterns between actors of the network. New visualization systems allow rapid exploration of such data with the help of interaction and data mining capabilities directly implemented in the interfaces. This coupling of visualization and data mining has been described as Visual Analytics and can help historians explore their networks with support of algorithmic support.

In this chapter, we present a general overview of the fields of SNA (**??**), HSNA (**??**), and Social Network Visualization (§2.4).

In this chapter, we first present a general overview of the field of visualization to get a first ideas of the utility of such techniques. Then, we present the social history discipline ans the use of quantitative methods in §2.2, before describing in depth how network analysis has been applied in the field in §2.3. Finally, we present in §2.4 how visualization and VA have been used in the context of HSNA, along the most popular systems currently used by social scientists.

2.1 Visualization

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

2.1.1 Information Visualization

Information Visualization focus on displaying abstract data to amplify cognition and gain insight on real world phenomena [?]. History is filled with classical examples of visual data displays which helped understand better specific events, such as Minard's map of Napoleon's march in Russia [?], or Snow's dot map of cholera cases in London which showed the proximity between street pumps and cholera infections [?]. If several examples of information visualization can be found thorough history, it mainly developed as a scientific field in the 1960s with Tukey's work on data analysis and visualization [?] and Bertin's publication of Semiology of graphics [?].

In this foundational work, Bertin described and organized the different visual elements usable in graphical information displays, and linked them to data features and relations types. An illustration of this work of categorization for network data is illustrated in Figure 2.1. Michael Friendly writes that "To some, this appeared to do for graphics what Mendeleev had done for the organization of the chemical elements" [?]. The development of computer science and the rise of hardware capabilities during the same time created a big need for data visualization. The amount of data stored increased exponentially [?] and descriptive statistics were not enough to understand the underlying structure of the amount and diversity of produced data. Visualization, leveraging the human visual system, allowed to rapidly see the hidden structure of a dataset and detect interesting and unexpected patterns very often unseen with classical statistical methods. One classical illustration of this is Anscombe's quartet [?] which consists of four datasets of points in $\mathbb{R}^2$ with the same statistical measures (mean, variance, correlation coefficient, etc.) but with very different structures, that plotting the data show immediately. The four datasets are illustrated in Figure 2.2.

A large number of visualization techniques emerged to make sense of the diversity 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

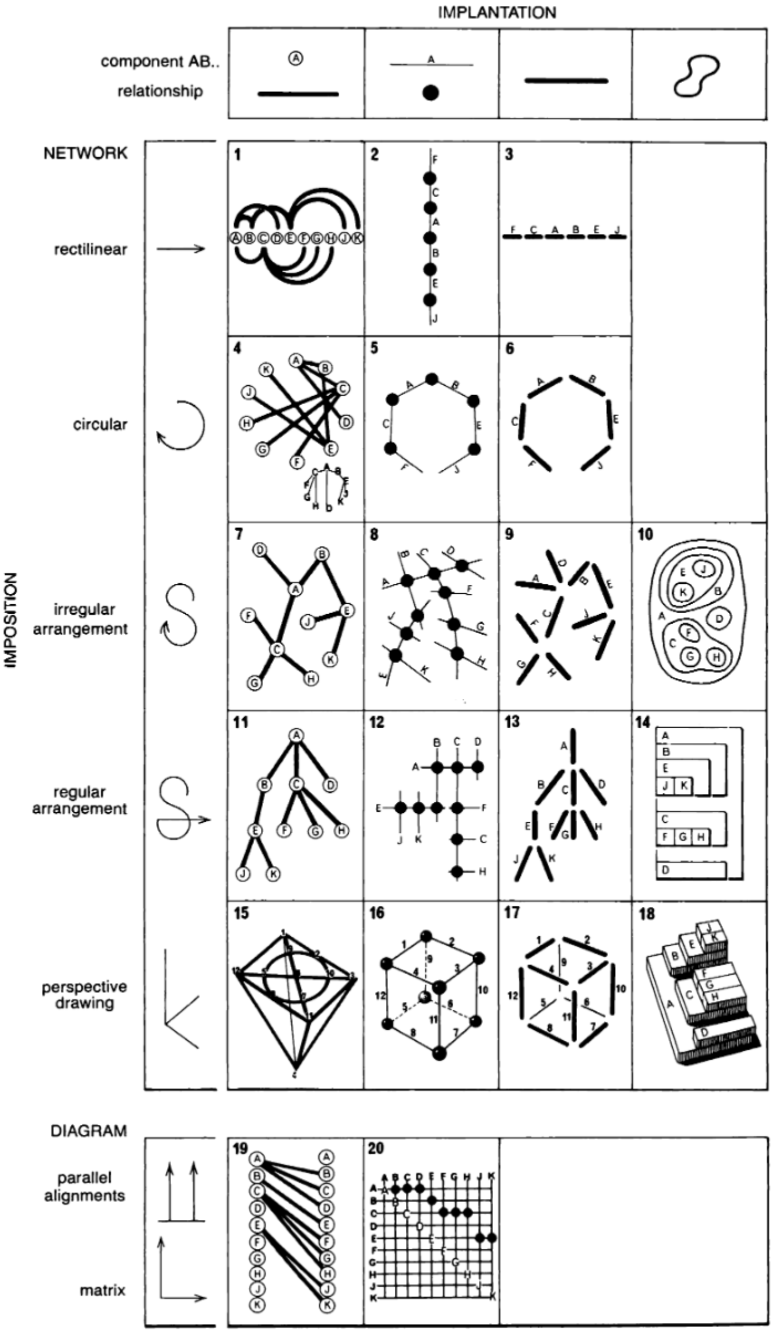


Figure 2.1 – Categorization of visual variables which can be used to represent network data, resulting in many different network representations. Image from [**?**].

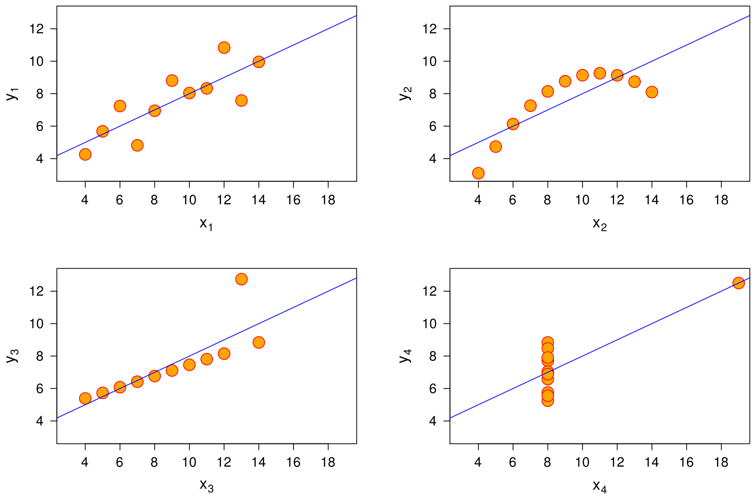
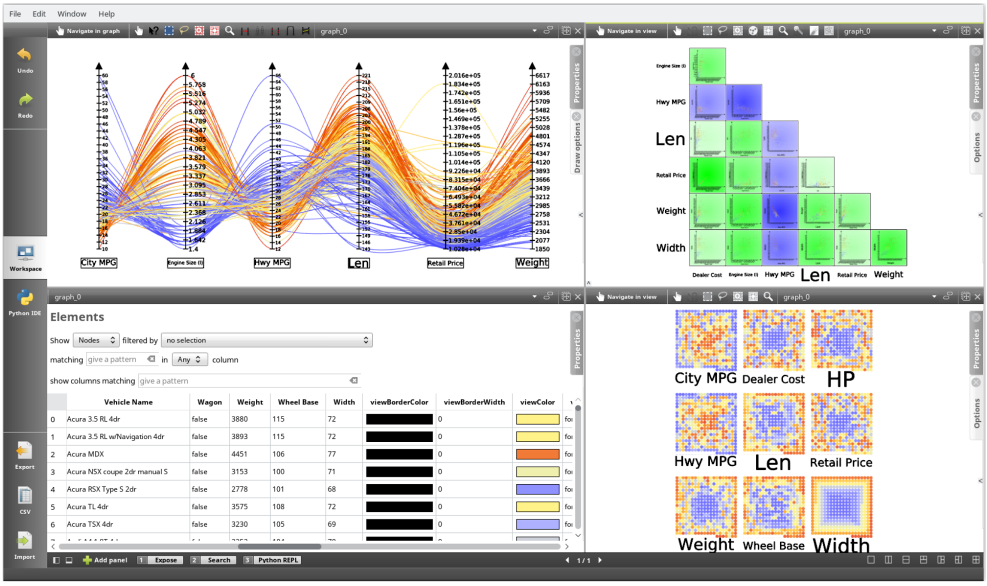


Figure 2.2 – Anscombe quartet. The four datasets have the same descriptive statistics (average, variance, correlation coefficient) but very different structures. Image from [**?**].

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 greater 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, first to scientific peers, and nowadays to broader audiences for example through the means of data journalism [?].

2.1.2 Visual Analytics

Visualization can also be used for exploratory aims, to gain new insights on the general structure of the data and potentially generate new hypotheses. This process has been characterized by Tukey in 1960 as *Exploratory Data Analysis* [?] and consist in trying to characterize the structure of a dataset with the help of visualization and statistical measurements. Visual exploration is enhanced by direct manipulation interfaces through interaction and usually follows the information-



Elements

Show Nodes filtered by no selection

matching give a pattern in Any column

show columns matching give a pattern

	Vehicle Name	Wagon	Weight	Wheel Base	Width	viewBorderColor	viewBorderWidth	viewColor
0	Acura 3.5 RL 4dr	false	3880	115	72	⬛	0	🟨
1	Acura 3.5 RL w/Navigation 4dr	false	3893	115	72	⬛	0	🟨
2	Acura MDX	false	4451	106	77	⬛	0	🟧
3	Acura NSX coupe 2dr manual S	false	3153	100	71	⬛	0	🟨
4	Acura RSX Type S 2dr	false	2778	101	68	⬛	0	🟦
5	Acura TL 4dr	false	3575	108	72	⬛	0	🟨
6	Acura TSX 4dr	false	3230	105	69	⬛	0	🟦

Figure 2.3 – TULIP software designed for application-independant network visual analytics [**?**]. The view shows a dataset among multiple interactive coordinated views. Users can also apply data mining algorithms on the data to extract interesting patterns.

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 pertinent 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 a very active research field. 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" [?].

VA consist in the generation of knowledge using visualizations and statistical models of the data, that the user can explore using interaction. Such systems have been developed in various empirical domains, such as biology, astronomy, engineering, and social sciences, to explore various data types such as multidimensional, temporal, geolocated, or relational (i.e., modeled into a network). Figure 2.3 shows the TULIP system, an example of a VA system developed for the analysis of network data. We discuss the uses of VA for HSNA in §2.4.2

2.2 Quantitative Social History

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

2.2.1 History, Social History, and Methodology

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

For Lucien fèbvre, philosophising was even constituting a "capital crime" [?, ?].

Retrospectively, we can still observe shifts in the objects of study of historians through time, and their relation to sources. History was at first mainly event-centered and was focusing in characterizing central figures of the past like rulers and artists or shed light on central events like wars or political crisis. This narrative approach to history has been criticized for its open interpretation of historical documents, which can introduce bias from the authors [?]. In the 1930s, March Bloch and Lucien fèbvre detached from traditional history by creating the "Annales school" (Ecole des Annales) which aimed at placing the human as a component of a broader sociological, political, and economic system with influences between each other [?]. They strongly advised to exhaustively search from archives, to ground historical results in documents, texts and numbers. This new way of studying past events and societies became successful in a profession in crisis, by bringing a new lens of study on various societal subjects more grounded in sources and with a better intelligibility. This school of thought can be seen as one of the biggest milestones for Social History, which focuses on the socio-economical aspects of societies and their changes through time, rather than an event-centric view of History. For example, in his thesis, Ernest Labrousse—a well known figure of Social History—tries to describe and explain the economic crisis of France at the end of the "Ancien Régime"[1] through the evolution of the economic power of different social groups such as farmers, workers, property owners etcinstead of solely describing memorable facts about the period [?]. Social History continued to evolve since the 1930s, introducing new methods and concepts, but always with the goals to describe periods and historical facts through a sociological lens and with a strong focus on sources and traceability.

2.2.2 Quantitative History

With the development of statistical methods and Computer Science, quantitative approaches of History emerged in the 1960s with the goal of analyzing numeric data directly extracted from historical documents. Economists led this first wave of quantification by studying past events using economical concepts and data. This approach, called "new economic history" or "cliometrics" was popularized by Fogel's study on the economic impact of the development of railroads in America [?] and Fogel and Engerman's controversial work on the economy of slavery [?]. In the latter study, they extracted numbers of a sample of 5000 bills of slave sales from New Orleans to support the controversial claims that slavery was economically viable and that slaves had a decent material life, which brought up heated debate among the scientific community and the broad audience [?]. These kinds of approaches rapidly started to be used in other related domains such as demography, social history, and political history, sometimes rebranded as

[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.

"new social history" and "new political history" [?] As extracting the data from raw documents and uploading it in computers—which were shared among whole departments—was very time-consuming at that time, "new history" projects often relied on a high division of labor among researchers, assistants, and students who operated with punch card operators [?]. Many saw the future of social sciences in computer programming, as Le Roy-Ladurie who wrote in 1968 "The historian of tomorrow will be a programmer, or he will not exist" [?].

However, quantitative methods started to be criticized in the 1980s with a vague of disillusionment, for several reasons. Stone was the first to raise his voice in 1979, after participating himself in several of those ambitious projects: "It is just those projects that have been the most lavishly funded, the most ambitious in the assembly of vast quantities of data by armies of paid researchers, the most scientifically processed by the very latest in computer technology, the most mathematically sophisticated in presentation, which have so far turned out to be the most disappointing" [?]. First, many researchers of this first wave dispensed themselves of source criticism, leading to simplification, anachronisms—such as using modern analytical categories and indices like the GDP—, and taking the numeric data from historical documents as objective. These problems could be in part explained by the fact that the work process was highly divided, meaning that the people analyzing the data did not necessarily inspect and read the original historical documents in depth. Secondly, the popularity of these methods made practitioners forget about the many biases inherent to statistics, such as the sampling bias, or the fact that historical data is essentially uncomplete data. This resulted in the computation of long data series and aggregates which were sometimes nonsensical given the gaps in the sources [?]. Finally, many historians raised their voice against the study of long-term trends instead of focusing on specific events and individuals. They challenged aggregations procedures and its assumptions, trying to go back to a more complex history by pointing that phenomena have to be studied and understood through several scales [?]. Indeed, computing correlations and aggregates at a national level greatly simplify complex phenomena, and misses specific group and individual related behaviours. Still, if their adoption remains slow and sometimes criticized among historians, quantitative methods provide tools to store, explore, and analyze historical documents systematically if used appropriately (i.e. not trying to bias the analysis, and not losing the trace of the original sources), especially that those methods highly evolved since the 1960s.

2.2.3 Digital Humanities

Digital Humanities is sometimes described as the second wave of computational social sciences [?]. The term has gained popularity since the 2010s and refer to "research and teaching taking place at the intersection of digital technologies and humanities. Digital Humanities aims to produce and use applications and models that make possible new kinds of teaching and research, both in the humanities and in computer science (and its allied technologies). Digital Humanities also studies the impact of these techniques on cultural heritage, memory institutions, libraries, archives and digital culture." [?]. If the first wave of computational social sciences focused a lot on statistical methods such as regression models, correlation testing, and descriptive measures (mean, median, and variance) to make conclusions, digital humanities focuses more on the use of digital tools for exploration, teaching, and communication of humanities datasets and concepts, leveraging design, infographics, and interactive systems [?]. In the context of historical research, the term Digital History have been coined as "an approach to examining and representing the past that works with the new communication technologies of the computer, the Internet network, and software systems. On one level, digital history is an open arena of scholarly production and communication, encompassing the development of new course materials and scholarly data collections. On another, it is a methodological approach framed by the hypertextual power of these technologies to make, define, query, and annotate associations in the human record of the past. To do digital history, then, is to create a framework, an ontology, through the technology for people to experience, read, and follow an argument about a historical problem." [?] Research which label itself as Digital History pivot around the curation and digitization of historical archives, the identification of historical concepts through computational and exploration methods, but also their communication to the general audience through digital technologies.

Many Digital History projects are thus multidisciplinary by essence and involve several teams of researchers, such as the Republic of Letters project which consisted in digitizing, storing, and exploring letters of scholars across the world, in a common hub and using shared visualization tools [?]. It resulted in the elaboration of curated datasets and visualizations concerning the correspondence of various scholars such as Voltaire, Benjamin Franklin (see Figure 2.4), or John Locke, accessible in the same place by researchers and the general audience. With modern technologies and infrastructures, it also becomes possible to study large historical databases—often labeled under the term "big data"—as with the Venice Time Machine project [?] which aims at digitizing and analyzing thousands of documents from the archives of Venice to understand the political, geographical, and sociological dynamics of the cities across generations and centuries. Yet, some historians raised concern about this type of project, fearing that it could rapidly bring the same type of issues that we saw during the first wave of quantification, especially for big projects involving many actors and high ambitious goals [?].

Many projects which claim themselves as Digital History also leverage new methods compared to the 1960s and 1970s, such as the use of network methods and concepts [?]. Examples are the Viral Texts [?] and Living with Machines [?] projects which respectively study nineteenth-century newspapers and the industrial revolution by translating their sources into analyzable networks. We discuss more in depth the related work of network analysis for historical research in **??**.

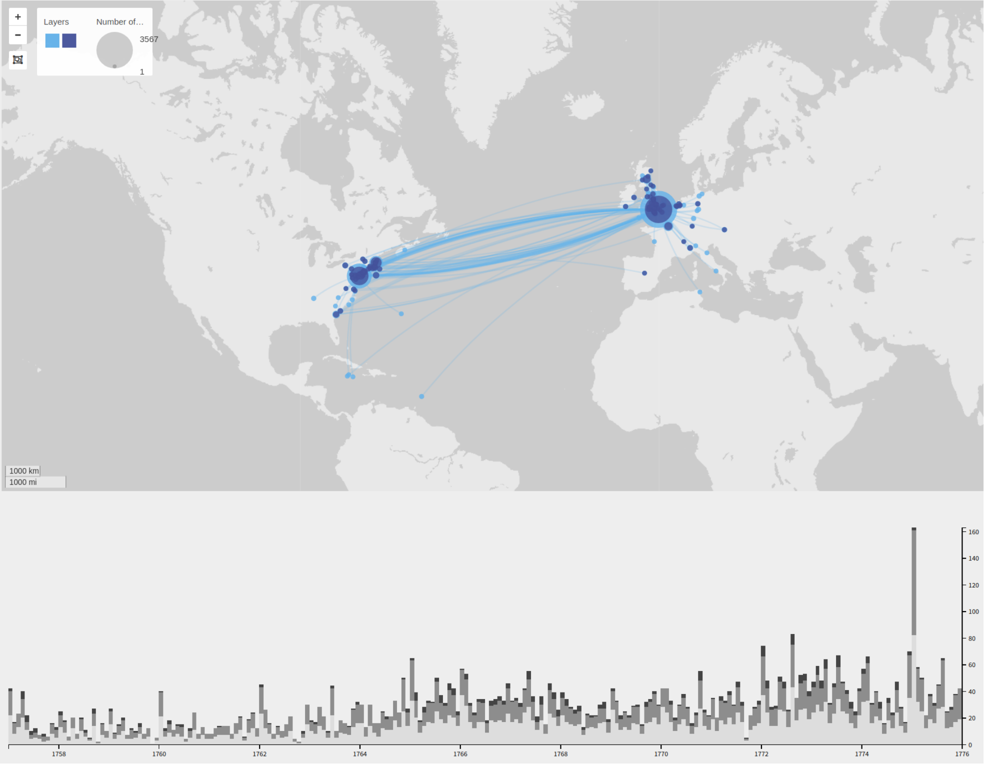


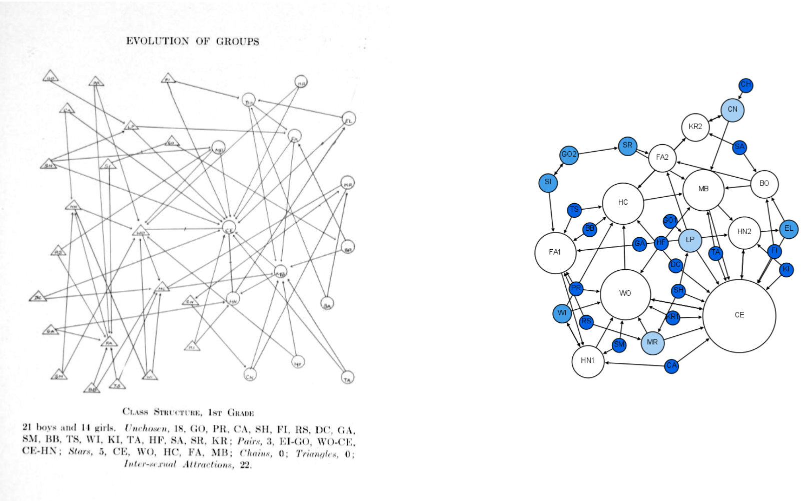
Figure 2.4 – Correspondence letters of Benjamin Franklin and his close relationships, using a map and an histogram, accessible online on the republic of letter website [**?**].

2.3 Historical Social Network Analysis

Historians started to use network analysis to study relational structures and phenomena of past societies in the 1980s, using similar methods developed by sociologist under the label of SNA. SNA is defined as an "approach grounded in the intuitive notion that the patterning of social ties in which actors are embedded has important consequences for those actors. Network analysts, then, seek to uncover various kinds of patterns. And they try to determine the conditions under which those patterns arise and to discover their consequences" [?]. the use of networks emerged in response to traditional sociology methods using pre-defined taxonomies and social categories to understand and explain sociological behaviors and phenomena, which could introduce bias. By modeling real observed social relationships and interactions with networks and by using mathematical and statistical methods to study those, sociologists have been able to explain sociological phenomena and describe sociological interactions through their direct observation modeled as networks. SNA is now a well-praised methodology in sociology and has been extended to historical research to study relational concepts such as kinship, friendships, and institutions of the past. Social historians leverage their documents to extract relationships between entities—often persons—that they model into networks. Leveraging network measures and visualization, 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 generalize on how social relations operate, and what are the social specificity of specific groups and individuals [?]. Traditional methods try to answer those questions using classical social classifications such as age, social status, profession, and gender, typically collected from surveys and interviews. Criticism pointed that this type of division is often partially biased and comes from predefined categories which are not always grounded in reality [?], and that using random sampling of individual with such methods remove them from their context. The sociologist Allen Barton wrote in this regard "For the last thirty years, empirical social research has been dominated by the sample survey. But as usually practiced, using random sampling of individuals, the survey is a sociological meatgrinder, tearing the individual from his social context and guaranteeing that nobody in the study interacts with anyone else in it" [?]. 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 socially structured [?]. He developed sociograms to visually show friendships between people with the help of circles representing persons and lines modeling friendships.



Class Structure, 1st Grade

21 boys and 14 girls. *Unchosen*, 18, GO, PR, CA, SH, FI, RS, DC, GA, SM, BB, TS, WI, KI, TA, HF, SA, SR, KR; *Pairs*, 3, EI-GO, WO-CE, CE-HN; *Stars*, 5, CE, WO, HC, FA, MB; *Chains*, 0; *Triangles*, 0; *Inter-sexual Attractions*, 22.

Figure 2.5 – Moreno's original sociogram of a class of first grades from [**?**] (left). The diagram shows 21 boys (triangles) and 14 girls (circles). The same sociogram plot using modern practices generated from Gephi from [**?**]. The color encodes the number of connections incoming.

Figure 2.5 shows one of Moreno's original sociograms to depict friendships in a class of first grades (left). Sociometry tremendously helped disseminate the metaphor of networks to model and understand social structures and phenomena. It was during the 1960s that sociologists and anthropologists took these concepts further and formalized SNA using graphs and mathematical methods, following the emergence of Graph Theory studies in the 1950s by Mathematicians such as Erdős [?]. Sociologists already had structural theories of social phenomena, and they rapidly saw the 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 [?]. The use of network measures let sociologists explain social phenomena through the formal description of real observations of relationships modeled as network.

2.3.2 Methods and Measures

Many measures and algorithms have been proposed in the network science and SNA literatures to characterize the structure of a network and relate it to social behaviours and phenomena. Networks measures are either global or local, which

[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.

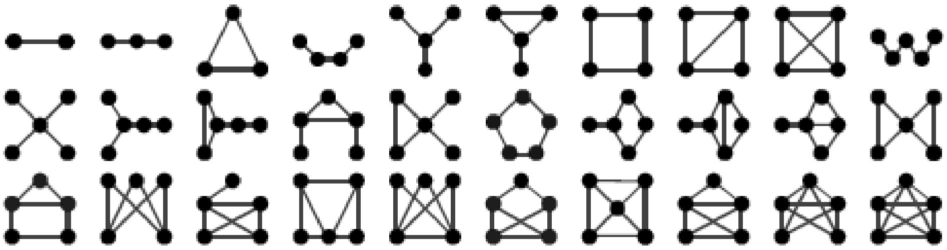


Figure 2.6 – All possible graphlets of size 2 to 5 for undirected graphs

allow to either make high level conclusions on the general structure of social relationships or individual behaviours. Widely used global measures include for example the density and the diameter, which give insight on the sparsity of the network and how distant on average are two random pairs of nodes. Conversely, local measures give information on the structural position of a node compared to the rest of the network. Centrality—probably the most used local measure—allows to formally compute a measure of how important or central are individuals in the network. As defining what an important node is ambiguous, several types of centrality have been proposed such as the degree, betweenness, and closeness centrality, which respectively measure the number of connexions, how nodes connects different groups, and how close are the nodes compared to the rest of the network.

More generally, sociologists aim at identifying recurring patterns of sociability between actors, and link it to other behaviours, measures, or qualitative knowledge. These patterns can for example be small unconnected components, cliques, or bow-ties structures. Groups of nodes similarly located (central or distant) and having similar shapes are sometimes referred as structurally equivalent [?]. Instead of observing complex shapes, network scientists have also been interested at studying relationships at the lowest possible scale, i.e., observing relations between sets of 2 and 3 nodes at one, also called dyads and triads [?]. This reflects on Simmel's formal sociology, where he already referred to dyads and triads as the primal form of sociability [?]. More recently, graphlet analysis extended this concept to every pattern of N-entities [?, ?].

Graphlets are defined as 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.

Figure 2.6 shows all graphlets of size 2 to 5, for undirected networks. Graphlets counting shows that graphlets are not found in a uniform distribution in social networks [?], thus revealing that social networks do not have the same structure that random networks. 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 [?].

However, networks concepts, measures, and algorithms have not been used only to study groups, organizations, and societies, but also to focus on separate specific individuals. Indeed, two distinct methodologies emerged through the history of SNA: the structuralists and the school of Manchester [?, ?, ?].

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

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

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

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.

2.3.3 Historical Social Network Analysis

History started to use concepts and methods from SNA in the 1980s [?] in response to quantitative history, and to develop historical approaches—like *Microstoria* [?]—that focus on the study of individuals and small 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 necessarily studying the relational aspect of these concepts. Network research allowed them to model those relational entities more thoroughly using networks, thus allowing them to make new observations that it was not possible to make without taking into account the relational structure of these entities [?]. Since then, HSNA—a term coined by C. Wetherell in 1998 [?]—has been applied by historians to study multiple types of relationships, like kinship [?], political mobilization [?], administrative and economic patronage [?]. If these approaches fall under some of the same critics as quantitative history [?] like leading of trivial conclusions, it still led to classical work and interesting discoveries, such as the study of the rise of the Medici family in Florence in the 15th century by Padgett & Ansell [?], or Alexander & Danowski study on Cicero's personal communications [?]. In this work, they modeled the communication of Cicero into a network using 280 letters written by him between 68 B.C. and 42 B.C. It allowed them to study the relationships between knights and senators—which is a subject of interest in Roman history—and concluded that knight-knight interactions were very rare compared to senator-senator and senator-knight interactions. Cicero communication network is illustrated in Figure 2.7.

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

2.3.4 Network Modeling

Constructing a network from historical documents, which can vary tremendously in their formats and structures, is not a trivial task [?]. The most straightforward and well-known approach consists in constructing a 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

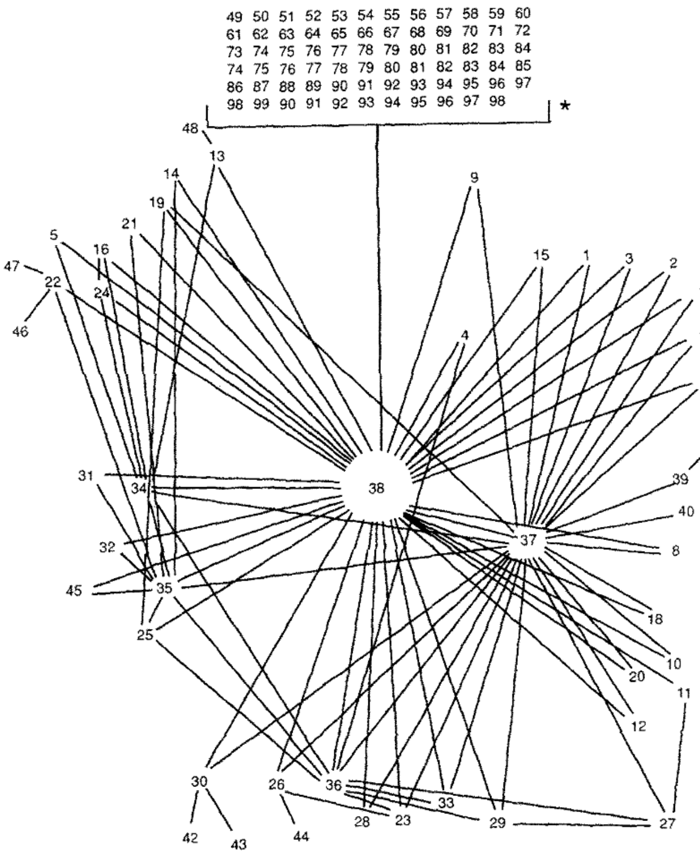


Figure 2.7 – Cicero personal communication network represented with a node-link diagram. Image from [**?**]

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 the importance of actors or relations with weighted networks, multiple relationships with multiplex networks, dynamics of relations with dynamic networks.

Weighted networks model the importance of relations, with a weight w attributed to each edge $e = (u, v, w)$, with $u, v \in V$, $e \in E$, and $w > 0$. Multiplex networks allow to model multiple kinds of relationships between actors, such as spouses and witnesses relations for an historical network constructed from marriage acts. In that case, each edge $e = (u, v, d)$ of the graph $G = (V, E, D)$ have a type $d \in D$ which characterize the relation. In the example of marriages, $D = \{spouse, witness\}$. Most relations extracted from historical documents also often contain time information, which can be modeled into dynamic networks. Many dynamic network models have been proposed [?], depending if the time is encoded in the nodes, the links, or both, and if entities have a timestamp of death. As it is often hard to infer the end of social relationships from the trace of historical documents, we only consider in this thesis models which give a timestamp to either nodes or edges, such that $G = (V, E, T)$ with vertices consisting of tuples (u, t) and edges of triples (u, v, t), with $t \in T$.

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 [?]. Formally, each node have a type $b \in B$, with $G = (V, E, B)$, $card(B) = 2$, and for each edge $e = (u, v) \in E$, the types b_u and b_v of u and v are not equal $b_u \neq b_v$. Many social situations or documents can be modeled in these terms (affiliation lists or co-authoring). Multivariate networks, i.e., graphs, where vertices and edges can be assigned multiple "properties" or "attributes", are less used in SNA. These attributes are often considered secondary, the emphasis of SNA being on the topology, its features, measures, and evolution.

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

When creating a network, sociologists and anthropologists can use direct observations of the real world, which is not the case for historians who only have access to biased and partial sources. Indeed, the documents historians inspect are often produced by the political and economical elite of the time, and include the subjective view of the authors, especially for literary sources (letters, journals, books, etc.). Historians therefore need to take a critical view on the sources by acknowledging the position of the authors of the documents compared to the rest of the society, and include it in the analysis [?]. Furthermore, the partiality of the sources often do not allow to have access to all possible relationships types of individuals. For example, if many formal relations can be extracted from official documents such as marriage acts and census, informal relations such as friendships can exist without leaving any written trace [?]. Even for official relationships such as parents and witnesses, there are high chances for missing documents, which do not allow to make too general and finite claims, such as "X is always the case" or "XX is never the case" [?]. Social historians therefore have to take into account the partiality and ambiguity of their sources into their analysis, in order to avoid including the bias inherent to their data into their high level historical conclusions.

2.4 Social Network Visualization

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

2.4.1 Graph Drawing

Sociologists rapidly saw the potential of graphically showing relationships between individuals, to better comprehend the underlying social structure and communicate their findings [?]. Moreno elaborated sociograms to visually show friendships among schoolchildren with circles and lines to respectively show children and friendships ties [?]. This type of representation—commonly called node-link diagram—is the most widely used in social sciences, as it is rapidly understandable and effective for small to medium-sized networks which are predominant in social sciences. The most used social network visual analytics software such as Gephi [?] and Pajek [?] are based on this type of representation and allow a fully integrated exploration and analysis with the help of various algorithms. Finding an optimal placement for the nodes is however not that simple as several metrics can be optimized depending on the desired drawing, such as the number of edge crossings, the variance of edge length, orthogonality of edges, etc [?, ?]. Figure 2.8 shows some of these metrics, synthesized by Kosara and al. [?]. In Figure 2.5 we can see the difference in readability between the original manual layout (left) and an automatic one (right). Automatic layouts which aim at optimizing readability metrics give clearer diagrams. The number of edge crossings is often considered the most important measure, but finding a drawing with the optimal number of crossings is an NP-Hard problem, meaning that heuristics are needed for most real-world use cases. A large number of algorithms have been designed such as force-directed ones, modeling the nodes as particles that repulse each other and are attracted together when connected with a link that can be seen as strings. Other visual techniques have been proposed to represent networks such as matrices, circular layouts, and arcs, but are less used in social sciences [?]. Still, Matrices have been shown to be more effective than node-link diagrams for several tasks such as finding cluster-related patterns, especially for medium to large networks [?, ?].

As social scientists are using more complex network models such as bipartite or temporal networks, more sophisticated representations are needed. The visualization community developed new representations to visualize other network types such as dynamic hypergraphs with PAOHVis [?], clustered graphs with NodeTrix [?] (illustrated in Figure 2.9), geolocated social networks with the Vistorian [?], and

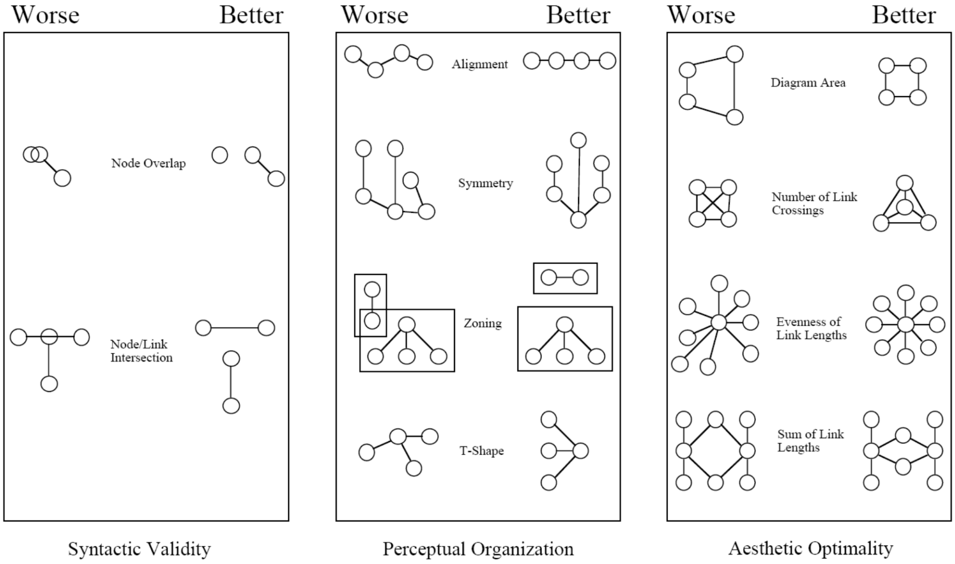


Figure 2.8 – Different criteria are proposed to enhance node-link diagram readability. Image from [**?**]

multivariate networks with Juniper [?]. However, these new network representations take time to be adopted by social scientists who rarely use those.

2.4.2 Social Network Visual Analytics

Social scientists use visualization and analytical tools to gain insight on the structure of their finalized network data. Most widely used tools are Gephi [?], Pajek [?], Ucinet [?], and NodeXl [?] which provide node-link diagrams, implementations of network measures, algorithms, and clustering capabilities. Other SNA visualization tools have been proposed in the past such as Visone [?]. However, those softwares often do not include interaction and direct manipulation mech-

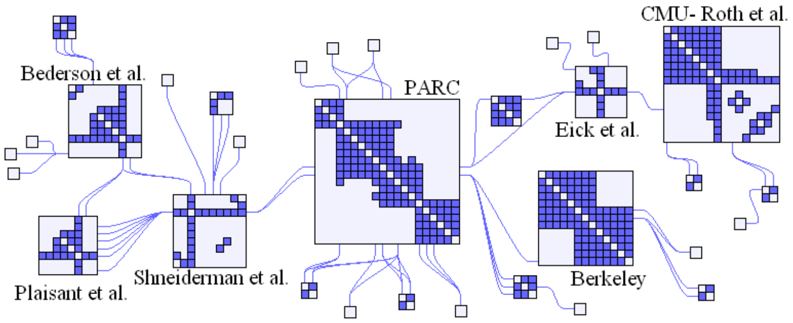


Figure 2.9 – NodeTrix system showing a scientific collaboration social network with clusters. Each cluster is represented as a matrix, Image from [**?**].

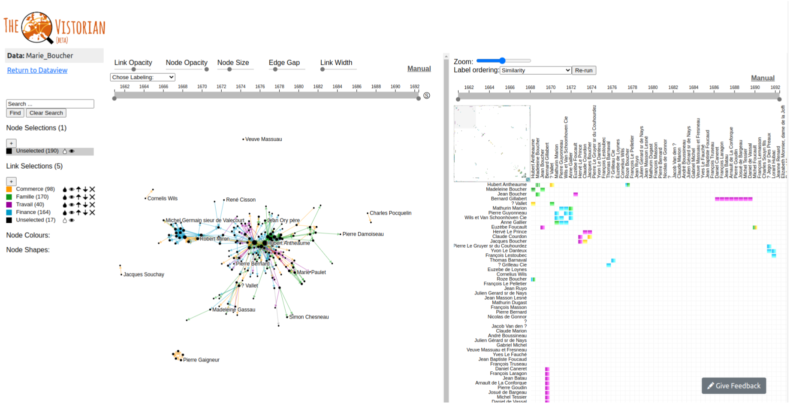


Figure 2.10 – Vistorian interface [?] used to explore a historical social network of business trades in the 17th century, with a coordinated node-link diagram and a matrice view.

	Representations	Network Measures	Clustering	Filtering	Direct Manipulation
Pajek [?]	■□□	■■■	■■□	□□□	□□□
Ucinet [?]	■□□	■■□	■□□	□□□	□□□
Gephi [?]	■□□	■□□	■□□	■□□	□□□
NodeXl [?]	■□□	■□□	■■□	■□□	□□□
Vistorian [?]	■■■	□□□	□□□	■■□	■■□

Table 2.1 – NodeTrix system showing a scientific collaboration social network with clusters. Each cluster is represented as a matrix, Image from [?].

anisms, making the analysis more tedious for social historians and pose usability issues. In contrast, the Vistorian [?] let social historians visualize their network with multiple coordinated views (node-link, matrice, arc-diagram, and map), filters and direct manipulation, but do not integrate analytical options. Figure 2.10 shows the Vistorian interface used to explore an historical social network.

I propose a classification of all those softwares in 2.1, which illustrate that most used software are analytical oriented (Pajek, Ucinet, Gephi, NodeXl) while the Vistorian is an interactive visualization tool. None of those software therefore fully correspond to the Visual Analytics label [?].

If analytical methods such as the computation of network measures, triad computation, or clustering provide a good framework to describe the structure of a network and link it to sociological explanation [?, ?], many social scientists such as historians are not trained in computer science and mathematical methods and therefore have trouble to first use those methods without guidance, but also interpret their results. This is particularly the case for black-box algorithms such as for clustering tasks: they usually end up trying several algorithms until they stumble upon a satisfactory enough solution [?].

Moreover, preparing and importing the data into visual and analytical software is complicated, as the annotation and network modeling process have not been globally formalized and every historian use different methods, formats and models. Many users do not succeed in importing their data in those systems without concrete help and guidance [?, ?] due to mismatches with data models, formats, or data inconsistencies (null values, white spaces, etc.). If they succeed visualizing their data, it often shows them these inconsistencies or errors such as duplications of entities or wrong attribute values. In other cases, they realize the network do not allow them to answer their sociological questions [?]. It leads to continuous back and forth between their analysis process inside the analysis tool they are using, and their annotation/modeling process, to correct errors or modify annotations. Interestingly, the network model choice plays a crucial role, as a simple network model representing only the persons (as it is often the case) makes it harder to trace back to the original documents containing the annotations from the network entities. Yet the majority of SNA systems enforce simple network models, making this retroactive process harder.

Some interfaces not primarily designed for social scientists incorporate data models encapsulating document representations, such as Jigsaw [?] which is a VA systems using textual documents as a data model, originally developed for intelligence analysis. It allows an analysis of the documents and their mentions of entities (persons, locations, institutions, etc.) through multiple coordinated views. Using such model allow to rapidly see errors and inconsistencies in the documents annotations that the user can directly correct, 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 the annotation, network modeling, correcting, and analysis steps, as errors and inconsistencies can cause high variations in the network structure and hence the analysis results [?].