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.

Table des matières

1 - Historical Network Analysis and Visualization

Social historians rely on textual historical documents to draw socio-economic conclusions about the past. They read and analyze all 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. During this process, they can use several methods developed in History to extract and analyze the information contained in the documents in a scientific 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 familiy, business or friendship ties—in a network, to be able to characterize and explain social behaviours through the description of the structure of the network. Historians following HSNA processes have inspired their workflow from Social Network Analysis (SNA), which is a well-known method in sociology and where a lot of methods and protocols had already been proposed when historians started to use similar approaches. Historians appropriated themselves this method and adjusted it to historical workflow which can vary from sociology as historians are limited in the documents they have and by their structure. They first have to annotate the documents to extract useful information, to then model it into an analyzable network. The annotation and modeling process is thus particularly complicated and specific to HSNA. Historians usually use social network visualization tools to confirm or generate new hypothesis once they successfully constructed their network. As the models used by historians are more and more complicated, new visualization systems are needed, first to analyze their networks, but also to help them in their HSNA process, from the acquisition of relevant documents to the final analysis and visualization steps.

1.1 . Social Network Analysis

The concept of SNA emerged in sociology in response to traditional methods using pre-defined taxonomies and social categories to understand and explain sociological behaviours 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 and manipulation. SNA is now a well praised methodology in sociology, which have also been appropriated by historians to study relational aspects of societies and institutions of the past.

1.1.1 . Sociometry to SNA

One of Sociology's main goal is to study social relationships between individuals and finding recurrent patterns and structures allowing to explain the behaviours of people and groups. Traditional methods try to explain social phenomena using classical social classifications such as the age, social status, profession and sex. For example, the social position of people living in a small city could be explained well by their age, demographics and social status which are traditional social categories. However, some criticism emerged that this type of division is often partially biased and come from predefined categories which are not always grounded in reality. Sociometry is considered as one of the basis 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 organization were functioning [51]. He elaborated sociograms as a way to visually show friendships between people with the help of circles representing persons and lines modeling friendships. 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 1950 by Mathematicians such as Erdos [17]. Sociologists already had structural theories of social phenomena, and they rapidly saw the potential of graphs 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 them in a sociology setting [7]. Using mathematical and network methods, it was possible to formally describe social relationships to make sociological conclusions grounded in real observations modeled as networks.

1.1.2 . Structuralism and Ego Studies

Lots of sociological studies used SNA concepts after it has been formalized. However, there was not yet strong protocols and methods to follow, and networks are an abstraction that can model different things in different ways. When looking retrospectively, we can see that two schools of thoughts emerged with different objectives and methods : the structuralists and the school of Manchester [19, 46, 20].

The structuralists are interested in observing the relational structures and patterns forming a network, to make parallels between them and the social behaviours of actors in real life [**?**]. They think the positions of the persons in the network and their relational patterns they are part of reflects 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 and business companies—and want to explain their functioning through the description of the internal shapes and

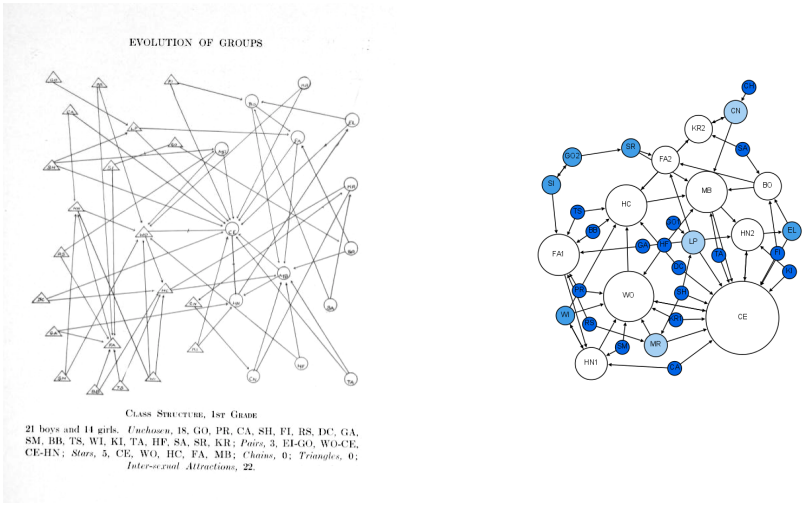


Figure 1.1 – Moreno original sociogram of a class of first grades from [50] (left). The diagram shows 21 boys (triangles) and 14 girls (circles). The same sociogram plot using modern practices generated from Gephi from [27]. The color encodes the number of connections incoming.

structures of the resulting networks. Thus, they try to construct networks which 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 focus on studying specific individuals and all their interactions in the different facets of their lives and in time. They typically want to explain certain behaviours and social characteristics of individuals by their relationships and interactions in all their complexity, and highlight the influence of some social aspects of one's life on other aspects. One famous example is Mayer's study on austral africa rural migrants going in cities [47] where he showed that the integration of urban mores and customs were directly correlated to the persons relationships networks in the city. Xhosa peoples still interacting with rural people of their village in the city were less changing their customs. This school of thought typically rely on the concept of ego network and more recently dynamic and multiplex networks. Ego networks are networks modeling all the direct relations of one central node—in this case a person—including the relations existing between the persons of this small network. They typically try to model the different types of relationships of a person, like their family, work and friendship ties and study them through time. By studying the ego network structure of someone, sociologists of this school try to leverage explanations on other social aspects of the persons like their social status, job and gender. It is also common to compare several ego networks to make correlations between the social relationships of individuals and other interesting

social categories.

These two methodologies of SNA are often not exclusives and current studies usually involve concepts and methods from both.

1.1.3 . Methods dans tools

Graph theorists and network scientists developed a myriad of measures and algorithms that sociologists appropriated themselves to describe and characterize social phenomena. When constructing networks, the first thing sociologists did was often to identify the main actors of the network, and explain why these actors were 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 the centrality have also been developed. Lots of types of centrality have been proposed, based on different criteria, as there are several ways of defining the more <emph>important</emph> 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 aimed 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 on low level relationships between people. This reflects on Simmel formal sociology, where he already referred to dyads and triads as primal form of sociability [65]. 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 an uniform distribution in social networks, thus revealing that these networks do not follow a random distribution. 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. In 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 as *communities*, and a lot of algorithms have been proposed to find these automatically.

1.2 . Historical Network Research

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 from the 1930s, with the rise of social history, by extracting information from historical textual documents and studying them with statistical methods in the 1960s. 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.

1.2.1 . Quantitative History

Traditionally, historians try to tell a story about protagonists and socio-economic facts in a given society by reading, understanding and linking together historical sources. This narrative approach to history has been criticized for its lack of traceability and the open interpretation of historical documents, which can introduce bias from the authors. To solve this problem, the "Annales school" (Ecole des Annales) proposed to characterize past social phenomena through the exhaustive and systematic analysis of historical documents [60]. Quantitative approaches then emerged in the 1960s with the appropriation of statistical and computer science methods to analyze data extracted from historical documents.

Using quantitative data, historians were able to make numeric conclusions on topics such as demography [32] 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 [28] 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 [37, 44]. Trying to understand complex historical phenomena is complicated and modeling the information contained in historical documents into quantitative datasets can rapidly simplify and distort the reality. Moreover, 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 [42]. Guildi and Aritage even criticize the decrease of interest of historians in working in archives [29]. 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 loosing 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 [36] 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.

1.2.2 . Historical Social Network Analsyis

History started to use concepts and methods from SNA in the 1980s [72] in order to criticize quantitative history concepts and results, and to develop historical approaches—like *Microstoria* [25]—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 to model those relational entities more thoroughly using networks concepts, thus allowing 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, similarly to SNA. Since then, HNR has been applied by sociologists and historians to study multiple kinds of relationships, like kinship and political mobilization [45], administrative and economic patronage [53], etc. If these approaches fall under similar critics of quantitative history [41] as for example the leading of trivial conclusions, it still led to classical works and interesting discoveries. One famous example is the study of the rise of the Medici family in Florence in the 15th century by Padgett [57], where he explained the rise of power of this family by their central position in the trading, marriage and banking networks of the powerful families of Florence. Figure 1.2 shows the different networks of Florence families where we can see the central position of the Medici. lots of historians are using and continuously improving the HNR method which can be very effective to study relational historical phenomena [39]. 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 formal sciences with their own practices [57, 59].

1.2.3 . Network Modeling

Constructing a network from historical documents, which can vary a lot in their formats and structures is not a trivial task. The most straightforward and well-known approach consists in constructing a social network based on a simple graph $G = (V, E)$ with V a set a 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. HNR network models have evolved over time to better take into account concrete properties of social networks, such as types of actors using labeled networks, the importance of actors or relations with weighted networks, mixed relationships with multiplex networks, dynamics of relations with dynamic networks. Bipartite networks have been proposed to model relations between two types of entities, such as organization and employees where the relations link employees to organizations but not employees to employees or organizations to organizations. Many social situations or documents can be modeled in these terms (affiliation lists or co-authoring). Multivariate networks, i.e., graphs, where vertices and edges can be assigned multiple "properties" or "attributes", are less used in SNA. These attributes are often considered secondary, the emphasis of SNA being on the topology, its features, measures, and evolution.

Historians, demographers, sociologists, and anthropologists have also been de-

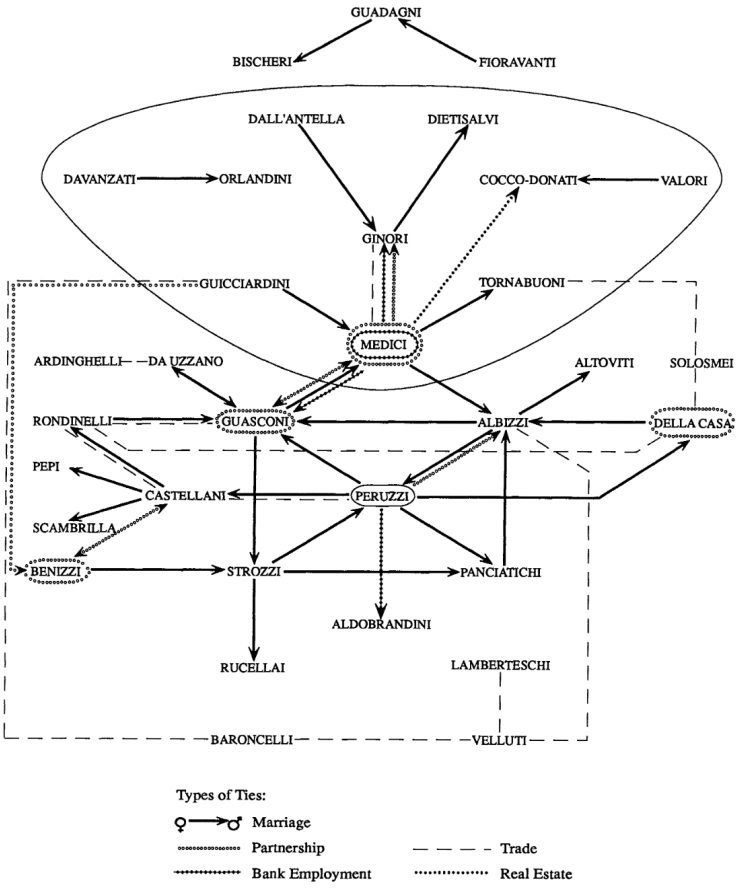


Figure 1.2 – Marriage, partnership. trading, banking and real estate networks of the powerful families of Florence from [57]. We can see the central position in the network of the Medici Family.

signing specific data models for their social networks, based on genealogy or more generally kinship [31]. For genealogy, the standard GEDCOM [23] 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 [30].

1.3 . Social Network Visualization

Practitioners of SNA and HNR have always depicted visually their networks for validation and communication purposes, mostly using node-link diagrams. With the increase of average network size 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 now following exploratory approaches using Visual Analytics (VA) tools, to describe more in depth their data and generate new interesting hypothesis, using interaction and exploration capabilities.

1.3.1 . Visualization

Data Visualization consists in graphically displaying data in the purpose of enhancing human cognition capabilities to understand and communicate ideas and phenomena. History is filled with classical examples of visual data displays which helped understand real phenomena, such as Minard's map of Napoleon march in Russia [21], or Snow's dot map of cholera cases in London which showed the proximity between street pumps and cholera infections [67]. If several examples of data 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 [69] and Bertin publication of Semiology of graphics [5]. 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. Friendly says that "To some, this appeared to do for graphics what Mendeleev had done for the organization of the chemical elements" [22]. 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 increase 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, allows 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 [3] which consists in 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 1.3.

Lots of visualization techniques emerged to make sense of the diversity of data produced, such as relational, temporal, spatial or network data. Subfields of Visualization emerged : **Scientific visualization** focus on visualizing continuous real data such as weather, spatial, and physics data, sometimes produced with simulations whereas **Information Visualization** is centered around the visualization of (multidimensional) discrete data points, often in an abstract way. **Visual Analytics** emerged later from Information Visualization by mixing data mining and more complex analysis process with traditional information visualization problematics. Historical Social Network visualization is closely related to Information Visualization and Visual Analytics, and good visualization systems for HNR use concepts and methodologies from those two fields.

1.3.2 . Social Network Visualization

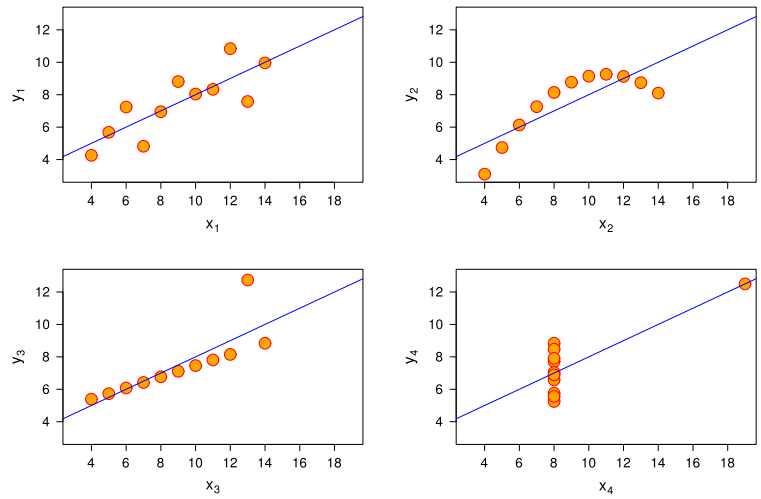


Figure 1.3 – Anscombe quartet. The four datasets have the same descriptive statistics (average, variance, correlation coefficient) but very different structures. Image from [3].

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 [50]. 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 [4] and Pajek [54] 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 number of edge crossings, the variance of edge length, orthogonality of edges etc [10, 40]. Figure 1.4 shows some of these metrics, synthesized by Kosara and al. [40]. In Figure 1.1 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 as the most important measure, but finding a drawing with the optimal number of crossing is a NP-Hard problem, meaning that heuristics are needed for most real world use cases. Lots of algorithms have been designed such as force-directed ones, modeling the nodes as particles which repulse each other and are attracted together when connected with a link which 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

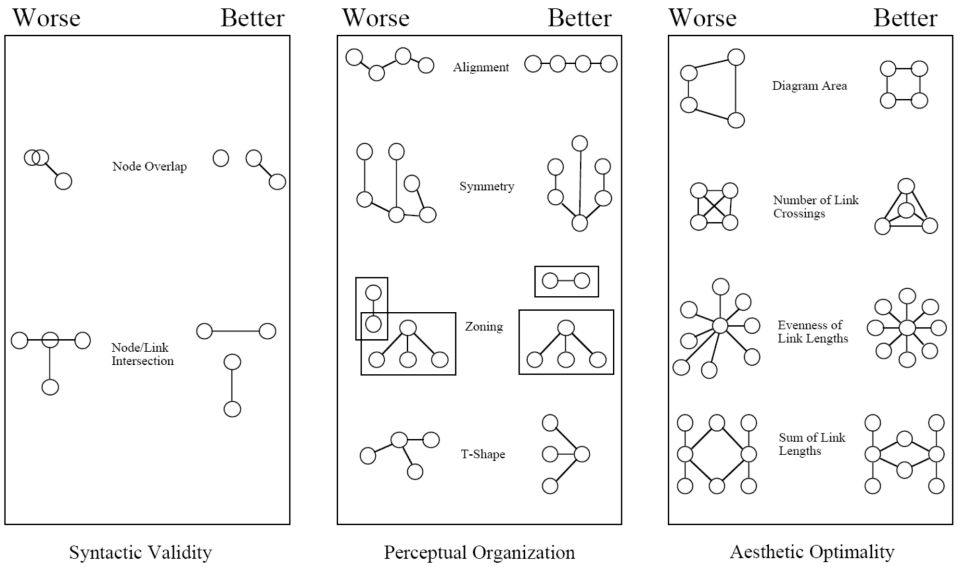


Figure 1.4 – Different criteria proposed to enhance node-link diagram readability. Image from [40]

in social sciences [49]. Still, Matrices have been shown to be better than node-link diagram for a lot of tasks such as finding cluster related patterns, especially for medium to large networks [24].

As social scientists started to use 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 [70], clustered graphs with NodeTrix [33] (illustrated in Figure 1.5), geolocated social networks with the Vistorian [64], and multivariate networks with Juniper [55]. However, these new networks representations take time to be adopted by social scientists who rarely use those.

1.3.3 . Social Network Visual Analytics

Social network visualization has mostly been used for confirmatory and communication purposes from its beginning. Social scientists often had hypothesis that they could rapidly verify by plotting the data. The same plots were often used for communication purposes, for example in a scientific paper or presentation. However, visualization can also be used for exploratory aims, to gain new insights on the data and potentially generate new hypothesis. This process has been characterized by Tukey in 1960 as *Exploratory Data Analysis (EDA)*. Exploration is mostly possible thanks to interaction, which allows to change the point of focus in the data to highlight interesting patterns, with the help of mechanisms like filtering, querying, sorting etc. As the average size of datasets keeps growing, exploratory tools are often needed to make sense of large datasets and generate interesting

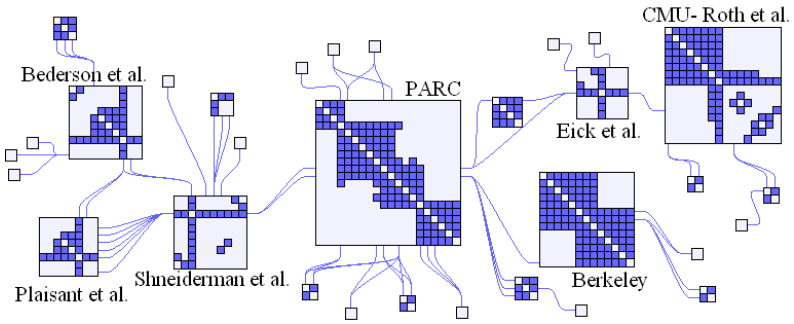


Figure 1.5 – NodeTrix system showing a scientific collaboration social network with clusters. Each cluster is represented as a matrix, Image from [33].

hypothesis.

Social scientists also often want to gain insight with the help of statistical and machine learning methods, that visualization only can not provide. More recent visual exploration interfaces incorporate automatic analytical tools along 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" [38]. Figure 1.6 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 generate and explore using interaction. Social scientists now frequently use VA systems to make sense of their data by using visualization, interaction, and data mining algorithms in their analysis loop to find interesting patterns and verify and create hypothesis. The most used social network VA tools are Gephi [4], Pajek [54] and NodeXl [66]. Figure 1.7 presents the Gephi interface showing a clustered social network, where each node is part of a cluster, encoded by color. They all let users visualize their networks with a node-link diagram, and allow an interactive exploration of the data with operations like filtering. Users can also analyze their data using network measures computed directly in the interface, and apply data mining algorithms such as clustering which results are explorable visually.

Unfortunately, social scientists are often not trained in computer science and mathematical methods, and a lot of them have been frustrated by VA tools and by how it was guiding their analysis in predefined ways. For example, lots of social network VA interfaces propose clustering features, allowing users to find interesting groups with the help of automatic algorithms, but social scientists often do not understand how the algorithms work and are not always satisfied with the

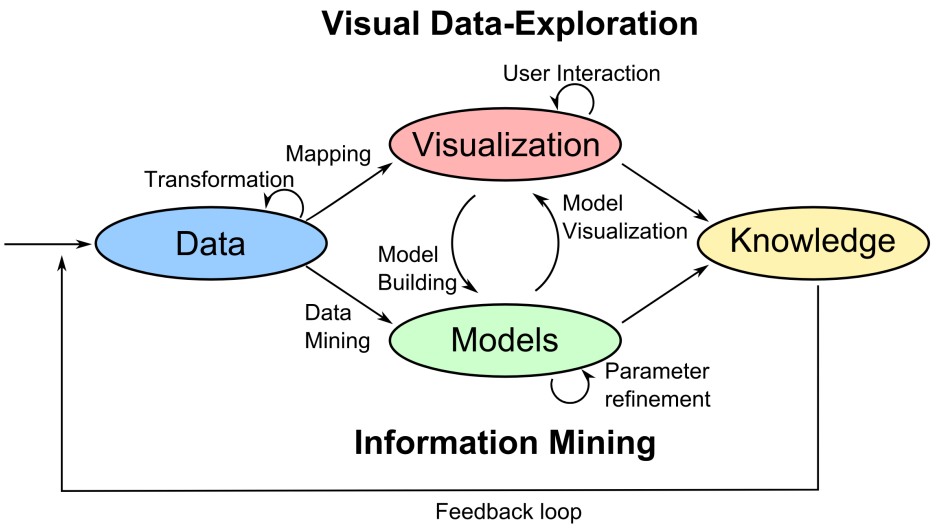


Figure 1.6 – Abstraction of the VA process. It is characterized by continuous interactions between the data, visualizations, models and knowledge. Image from [38].

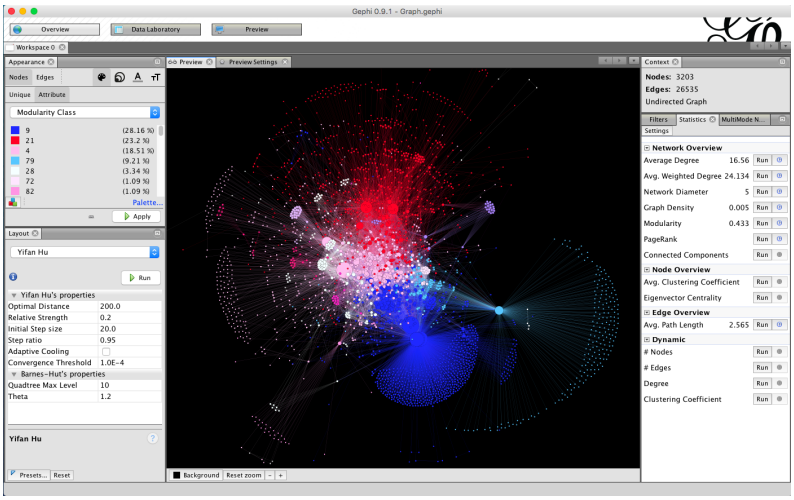


Figure 1.7 – Gephi [4] interface. The network is represented with a node-link diagram. Users can interact on the visualization and encode node and links visual attribute (color, size etc.) with network measures computed directly in the interface, such as the node degree, or clustering results.

results, as they can have knowledge from other sources not modeled inside the network. 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 and 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 do 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 chosen network model plays a big role on this 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 interface enforce simple network models, making this retroactive process harder. Some interfaces still incorporate data models encapsulating document representations, such as Jigsaw [68] which allows an exploration of textual documents with their mentioned entities. 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 helping 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.

2 - HSNA Process and Network Modeling

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 needs to generate many networks from the same documents/sources to visualize and analyze them. In this article, after describing and characterizing the workflow of Historical Social Network Analysis [72] (HSNA) 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*.

Social historians' goal is to characterize socio-economic phenomena and their dynamics in a restricted period and place of interest, and 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 Social Network Analysis (SNA), a method—sometimes referred to as a paradigm—which consists in modeling the social relationships between a set of entities—usually individuals—into a network. Much work has been done to adapt SNA to the context of historical document exploitation, and although several approaches coexist they can be brought together under the banner of Historical Social Network Analysis [72] (HSNA) or Historical Network Research [39] (HNR). When following an HSNA, historians collect documents, annotate them, and 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 different 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. This

paper proposes 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 make the communications of findings more replicable and transparent.

2.1 . Related Work

Since we already elaborated on the related work of SNA, HNR, network modeling and social network visualization in chapter 1, 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 of 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" [26] 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 [37, 43]. 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, hairballs correlations, statistical distributions, statistical testing, time series etc. [35, 42]. Similarly, the use of HNR for historical purposes increased in the recent years, and a lot or resources exist on how to use network methods and measures for historical research [41, 39].

However, few work has been done on describing and formalizing the process before the analysis part for a quantitative and network research workflow. Indeed,

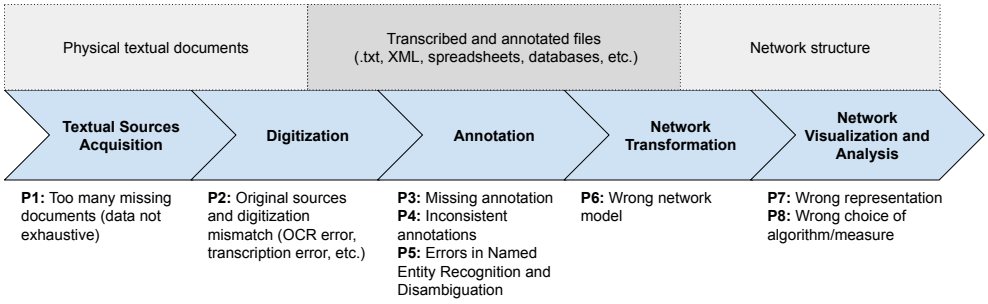


Figure 2.1 – HSNA workflow split in five steps : textual sources acquisition, digitization, annotation, network creation, network visualization and analysis. We list potential pitfalls for each step.

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 advices on how to annotate historical documents in the aim of using quantitative methods [37] 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* [15]. 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 [9]. Historians should be aware of this and chose a network model which fits their analytical goals.

2.2 . Historical Social Network Analysis Workflow

From the literature and discussions with historians' collaborators, 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 2.1 along potential and recurrent pitfalls.

2.2.1 . 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 docu-

ments 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)**.

2.2.2 . 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.

2.2.3 . 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 [13], 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) [8] 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 [16, 64]. 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.

2.2.4 . 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.

2.2.5 . 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 [20, 72]. 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 [63]. Similarly, social scientists can use data mining algorithms to highlight interesting and potentially hidden structure in the network, e.g. by using clustering algorithms revealing group structures [6]. **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 time a node appear 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 [56] 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.

2.3 . 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 [15]. 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.

2.3.1 . 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 [72] :** 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 [62] :** 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 [51]. However, historical documents rarely mention only one type of relationship and this model is thereby very limiting for HSNA.

— **Multiplex Unipartite Networks [18] :** 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 to model more complex social relations where people can have various social ties e.g. as parent, 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 [30] :** 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 analysis in HSNA encode the *roles* of the persons in the documents as link types [11]. 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 meaning 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 [48] :** in these networks, each node (vertex) is associated with a *layer* l and becomes a pair (v, l), allowing to connect vertices inside

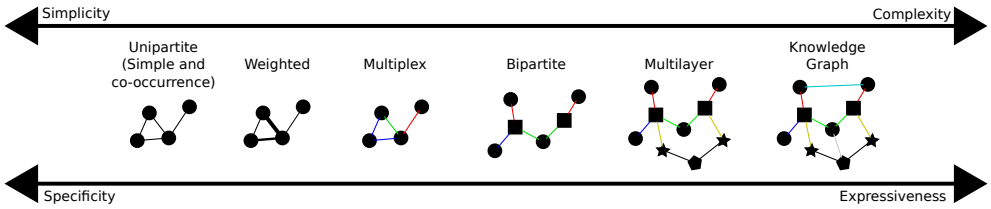


Figure 2.2 – Schematic representations of Different network models used for analyzing historical documents, ordered by complexity and expressiveness

a layer or between layers. These advanced networks have received attention from sociologists [12] and historians [71], 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)[34] :** 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 2.2 shows a schematic representation of the different network models. We can rank the models given two axes : simplicity/complexity and specificity/expressiveness. Currently, historians mostly construct unipartite networks (simple, co-occurrence, and weighted) which are simple and allow them to answer specific questions. However, those models 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 [41]. Several interpretations may coexist to explain why someone is central in the resulting network, which may be impossible to validate without encoding more information—such as the types of relationships—in the model. Depending on the schema of the annotations, it may be impossible to create more complicated networks at this step without redoing the annotation process which is costly in time and resources. On the contrary, too complicated models such as KG are difficult to create from the sources and are hard to visualize and analyze, especially for social scientists who are not trained in those kinds of formalisms. Using this kind of model usually require learning com-

plex query language to manipulate the data, such as the SPARQL language for KG. Therefore, we argue that historians should aim to model a network that is simple enough to manipulate, can be traced back to the original sources, and model well the social reality of the documents—i.e. having those three properties : *simplicity*, *traceability*, and *reality*.

2.3.2 . 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 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[11, 1].** 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 guaranty 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 types, and consist in *children, father, mother* for the birth events, *deceased* for the death event, *spouse* and *witnesses* for the marriages, and *family member* for the census events.
3. Analysis of migrations from Spain to Argentina through the **marriage acts at Buenos Aires in the 17–19th centuries [52, 61].** 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) [14].** 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 date 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 2.1.

As shown by Cristofoli [9], 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 to produce 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 relation makes 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 adds ambiguity in retrospect of the original documents, as it becomes impossible to trace back one link to one specific documents, as the same link could potentially refer to several ones [9].

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 an information like a date that we model as an attribute, we can store it as a document node attribute using a bipatite 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 2.1 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.

2.3.3 . Bipartite Multivariate Dynamic Social Network

We argue that bipartite multivariate dynamic networks verify the *simple traceability* and *reality* principles and model well the majority of historical documents. It has 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 [68] 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

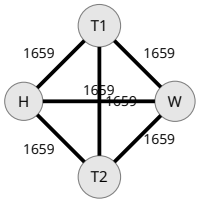
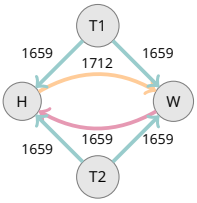
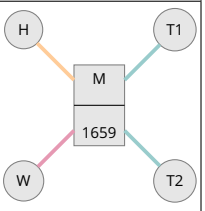
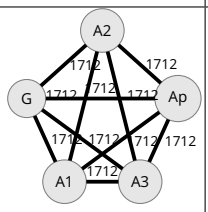
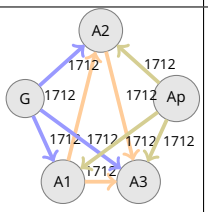
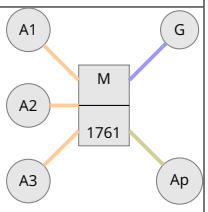
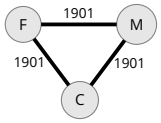
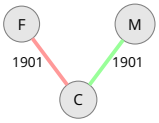
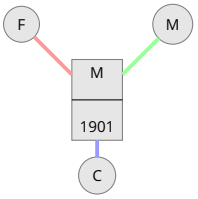
Original Document	Co-occurrence	Unipartite representation	Bipartite
20-4-1659 : Capitán Alonso MUÑOZ de GADEA , con Da. Francisca CABRAL LEAL de AYALA . Ts. : Agustín Gayoso , y Juan Guerrero. Al margen : "fue Oficial Real" , (f. 9v). Husband Wife Witness			
1712 : Construction of a church in Torino. 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 2.1 – Resulting networks using different models produced by one document of the examples detailed in subsection 2.3.2 : 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.

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 but 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 are applied.

2.4 . Applications

Several tools have been designed for visualizing dynamic bipartite networks that can also be considered dynamic hypergraphs [70, 58], 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[68], as it would allow to trace 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[64] now let users modify and clean their data in a table format if they see errors or inconsistencies.

2.5 . 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 document 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 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 in Figure 2.3 for the two cases where marriage acts mention birth information, and the case where only marriage related information is present in the document.

2.6 . 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

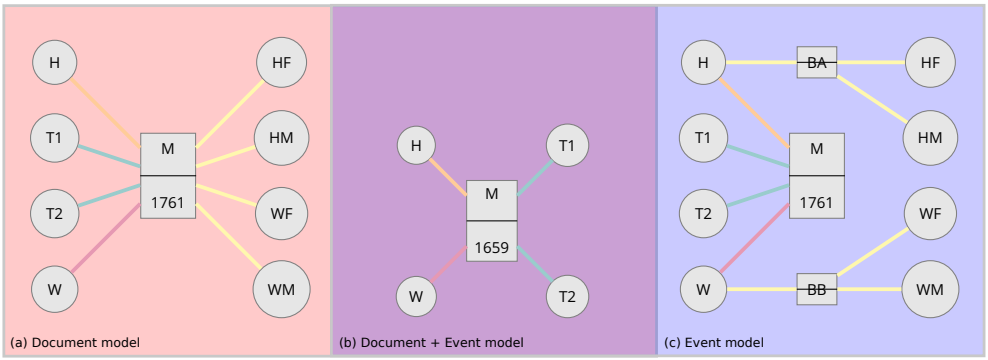


Figure 2.3 – bipartite multivariate dynamic network modeling for two cases of marriage acts of example #3. Some marriage acts mention the parents of the spouses, which is a relationships different than the marriage in itself. This case can be modeled using a document model (a) or an event model (c) by splitting the document into several different event nodes. The other case refer to document which do not mention the parents (b) and in that case the network represent 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.

characterize sociological phenomena using a common model and visual representation.

Bibliographie

[1] Mobilité et conflits. Travailler sur les chantiers de construction piémontais dans la première moitié du XVIIIe siècle. Coll. Histoire et Civilisations. Presses universitaires du Septentrion, Villeneuve d'Ascq, 2018.

[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 Trans. Vis. Comput. Graphics*, 27(2), February 2023.

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

[4] 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' 2009*. The AAAI Press, 2009.

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

[6] 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.

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

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

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

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

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

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

[22] 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.

[23] GEDCOM : The genealogy data standard.

[24] 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.

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

[26] 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.

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

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

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

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

[40] 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.

[41] Claire Lemercier. 12. Formal network methods in history : Why and how ? In Georg Fertig, editor, *Social Networks, Political Institutions, and Rural So-*

cieties, volume 11, pages 281–310. Brepols Publishers, Turnhout, January 2015.

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

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

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

[45] 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.

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

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

[48] 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.

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

[50] 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.

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

[52] 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.

[53] 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.

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

[55] 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.

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

[57] 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.

[58] 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.

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

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

[61] 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.

[62] 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.

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

[64] 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.

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

[66] 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 *Proceedings of the Fourth International Conference on Communities and Technologies*, C&T '09, pages 255–264, New York, NY, USA, June 2009. Association for Computing Machinery.

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

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

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

[70] 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.

[71] 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.

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