

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
Prénom Nom
Titre, Affiliation

Président ou Présidente
Rapporteur & Examinateur / trice
Rapporteur & Examinateur / trice
Examinateur ou Examinatrice
Examinateur ou Examinatrice
Directeur ou Directrice de thèse

Titre : titre (en français).....

Mots clés : 3 à 6 mots clefs (version en français)

Résumé : Lorem ipsum dolor sit amet, consectetur 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. Do-

nec 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, consectetur 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. Do-

nec 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	Introduction	5
1.1	Social History and Historical Social Network Analysis	6
1.2	Visualization and Visual Analytics	7
1.3	Historical Social Networks Visual Analytics	8
1.4	Contribution and research statement	8
2	Historical Network Analysis and Visualization	11
2.1	Social Network Analysis	11
2.1.1	Sociometry to SNA	12
2.1.2	Structuralism and Ego Studies	12
2.1.3	Methods and tools	14
2.2	Historical Network Research	15
2.2.1	History, Social History and Methodology	15
2.2.2	Quantitative History	16
2.2.3	Historical Social Network Analysis	17
2.2.4	Network Modeling	17
2.3	Social Network Visualization	19
2.3.1	Visualization	19
2.3.2	Social Network Visualization	20
2.3.3	Social Network Visual Analytics	21
3	HSNA Process and Network Modeling	25
3.1	Context	25
3.2	Related Work	26
3.3	Historical Social Network Analysis Workflow	27
3.3.1	Textual Sources Acquisition	28
3.3.2	Digitization	28
3.3.3	Annotation	29
3.3.4	Network Creation	29
3.3.5	Network Analysis and Visualization	30
3.4	Network modeling and analysis	31
3.4.1	Network Models	31
3.4.2	Bipartite Multivariate Dynamic Social Network	33
3.4.3	Examples	34
3.5	Applications	35
3.6	Discussion	37
3.7	Conclusion	38

4 ComBiNet : Visual Query and Comparison of Bipartite Dynamic Multivariate Networks with Roles	39
4.1 Context	39
4.2 Related Work	41
4.2.1 Graphlet Analysis	41
4.2.2 Visual Graph Querying	42
4.2.3 Visual Graph Comparison	42
4.2.4 Provenance	43
4.3 Task Analysis and Design Process	43
4.3.1 Use Cases	43
4.3.2 Tasks Analysis	46
4.4 The ComBiNet System	47
4.4.1 Visualizations	48
4.4.2 Query Panel	50
4.4.3 Comparison	57
4.4.4 Implementation	58
4.5 Use Cases	59
4.5.1 Construction sites in Piedmont (#1)	59
4.5.2 French Genealogy (#2)	60
4.5.3 Sociology thesis in France	61
4.6 Formative Usability Study	64
4.6.1 Feedback	64
4.7 Discussion	64
4.8 Conclusion and Future Work	65
5 PK-Clustering	67
5.1 Context	67
5.2 Related Work	70
5.2.1 Graph Clustering	70
5.2.2 Semi-supervised Clustering	71
5.2.3 Mixed-Initiative Systems and Interactive Clustering	71
5.2.4 Groups in Network Visualization	72
5.2.5 Ensemble Clustering	72
5.2.6 Summary	73
5.3 PK-clustering	73
5.3.1 Overview	73
5.3.2 Specification of Prior Knowledge	75
5.3.3 Running the Clustering Algorithms	75
5.3.4 Matching Clustering Results and Prior Knowledge	76
5.3.5 Ranking the Algorithms	77
5.3.6 Reviewing the Ranked List of Algorithms	78
5.3.7 Reviewing and Consolidating Final Results	79
5.3.8 Wrapping up and Reporting Results	84

5.4	Case studies	84
5.4.1	Marie Boucher Social Network	84
5.4.2	Lineages at VAST	85
5.4.3	Feedback from practitioners	87
5.5	Discussion	89
5.5.1	Limitations	89
5.5.2	Performance	90
5.6	Conclusion	90
6	Conclusion	93
6.1	Summary	93
6.2	Discussion	93
6.3	Perspectives	95
6.4	Conclusion	96

1 - Introduction

Social scientists such as historians and sociologists want to make sense of the structure and evolutions of the social relationships between people of a given place and time. Social Network Analysis (SNA), also called Historical Social Network Analysis (HSNA) when applied to historical data, is one of the main methods—sometimes referred to as a paradigm—to achieve this goal, and grew in popularity in recent years as the concept of networks became a widespread abstraction in many empirical sciences. It consists in modeling the social relationships between agents—such as persons and organizations—as a network and studying its local and global structure to make sociological conclusions, based on real observations. Compared to other more classical sociological approaches, “it is based on an assumption of the importance of relationships among interacting units. The social network perspective encompasses theories, models, and applications that are expressed in terms of relational concepts or processes.” [?]. Social historians rely on textual documents—such as marriage acts, birth certificate, and migration acts—and try to exhaustively extract all mentions of social relationships between persons that they model into a social network, where the nodes usually model persons, and links social relationships. Using mathematical computations and measures to characterize the network’s structure, they are able to make sociological conclusions on relational sociological phenomena based on real observations.

They also rely on Social Network Visualization (SNV) to make useful representations of social networks and are starting to use Visual Analytics (VA) tools that allow them an effective exploration and analysis of this type of data, based on direct interaction and manipulation to find relevant patterns and trends. However, the process leading to the analyzable network from the historical documents is quite complex and not well documented. Specifically, historians can model their networks in many ways, and choosing a model loosely can lead to simplifications and distortions in the social reality and the final analyses. Annotating the documents and cleaning the extracted information is also not trivial and can lead to errors and biases. Most VA tools currently focus on the analysis part only of the process and do not help social historians in their annotation, cleaning, and modeling steps. Moreover, automatic algorithms capabilities such as clustering are often available in SNA tools to help users make sense of the network structure, but interpreting the results from a sociological perspective is complicated especially without guidance, and social scientists have become frustrated by the use of such software [43]. The goals of this thesis are therefore first to characterize the HSNA process of social historians, and from characterizing recurring pitfalls, see how VA can help them in this complex process. Particularly, we study how VA can help them follow complex analyses to answer their sociological questions while keeping a high level of simplicity, traceability, usability, and with the right level of control.

1.1 . Social History and Historical Social Network Analysis

HSNA is now a widely used method among historians to study relational phenomena like kinship, business, and institutions of the past, using network concepts and methods. It appeared as an effective way to study their sources after a long process of methodological shifts and discussions [9]. We can trace it back to the birth of Social History with the “Annales School” in the 1930s, where Historians gained interest in socio-economic questions, and started to rely heavily on the exhaustive extraction and analysis of historical documents coming from archives [?]. Beforehand, History was mainly political and event-centered, as the majority of works consisted in narrating and characterizing specific events—such as wars and diplomatic alliances—while eliciting their causes and consequences, and describing the lives of historic figures, such as sovereigns. Social History shifted the focus by trying to link together sociological, economical, and political issues and by placing individuals at the center of these questions [60]. Later on in the 1960s, with the development of Computer Science, historians started to use quantitative methods to analyze data extracted from historical documents and make conclusions grounded in statistical results, in various subjects like demographics [32] or economics [?]. Around the same time, the use and study of networks started to become popular in various disciplines to study real-world relational phenomena based on mathematical computations and measures. A network is an abstraction based on graph theory concepts used to model phenomena based on relationships (called links) between entities (called nodes).

Sociologists started to use this concept to model social ties between agents of interest—very often persons—and study social phenomena through the description of the network structure, using the SNA methodology [?]. It allowed them to leverage quantitative measures from the network to make sociological conclusions based on real observations, instead of relying on traditional social categories such as age, job, and gender [?]. It has been applied to study various subjects like families, political institutions, schools, friendships, work environments, and sports clubs, with promising results. This method grew in popularity in recent years and has started to be used and formalized by historians in the 1980s, under the term Historical Network Research (HNR) or Historical Social Network Analysis (HSNA) [72]. It allowed historians to study relational phenomena of the past—such as families and institutions —through real social relationships directly extracted from historical documents and modeled into a network. These documents can be anything mentioning social relationships between actors of interest, such as marriage acts, birth certificates, census, migration acts, and work contracts. After selecting a corpus of documents (often extracted from archives), they manually annotate each document to extract the mentions of persons along their relationships to model those with a network. This is a long and tedious process that can result in small to large networks that they analyze using mathematical measures to make conclusions on historical phenomena.

Typically, social historians using quantitative and network methods work in a descriptive framework and leverage values and measures to make conclusions. Lemercier and Zalc say “Although history is not an exact science, counting, comparing, classifying, and modeling are nevertheless useful methods for measuring our degree of doubt or certainty, making our hypotheses explicit, and evaluating the influence of a phenomenon.” [42] Social historians, therefore, have hypotheses about their subject of study, that they can back up or refute with the help of quantitative and network results, similarly to the competing hypotheses workflow of Intelligence Analysis [?]. By pointing to evidence supporting or refuting hypotheses, they can give insight into the level of the plausibility of different claims. To verify their findings, historians rely on Social Network Visualization (SNV) to represent visually their data. It lets them find patterns to confirm or deny hypotheses and communicate their finding.

1.2 . Visualization and Visual Analytics

Visualization is the process of displaying data visually to leverage the human visual system and enhance cognition to gain insight into the data [?]. Using visual abstractions (such as size, color, and position) to display real data allows us to rapidly detect potential hidden structures and patterns. As data keeps growing in size with time, visualization is a powerful tool to gain insight into the underlying structure of various datasets. Visualization is mainly used for confirmatory and communication purposes, particularly in social sciences. Analysts generally have hypotheses that they want to test : they collect data, model it, compute measures, and plot the data visually to verify that it matches the results. They typically use similar plots to communicate findings, for example in scientific reports and presentations. Social scientists following a network analysis typically use node-link diagrams, which allow them to represent a network with circles and lines to respectively represent nodes and links. It let them confirm or deny some of their hypotheses by observing the network structure and communicating their results.

However, visualization can also be used for exploration, which can help to understand the underlying structure of data and generate new hypotheses. Tukey defined this process as Exploratory Data Analysis (EDA) in the 1960s [?], as a procedure to gain insight into the structure of the data by identifying outliers, trends, and patterns with the usage of visualization and statistical measures. Interaction in visualization helps this process by letting users navigate more easily on multivariate datasets with the help of direct manipulation of the data. The detection of patterns and trends can also be facilitated with automatic methods coming from data mining and machine learning fields, directly implemented in the visual analysis loop. This coupling of visual exploration and automatic data mining algorithms has been coined as Visual Analytics (VA) and is defined as the process of using interactive visualizations, transformations, and models of the data in an analysis workflow to

create knowledge [38].

EDA and VA are currently not widely used in HSNA, as social historians always have prior hypotheses before getting their data in shape. However, datasets that historians produce are getting bigger and bigger and VA tools are efficient in rapidly generating insights and new hypotheses.

1.3 . Historical Social Networks Visual Analytics

Currently, social historians spend a lot of time in their data acquisition, processing, and cleaning steps which lead them to the construction of a network. They typically visualize and analyze their network at the very end of this process only. This can be explained in part because most VA tools focus only on this final analysis and not on the preprocessing steps. Social historians collect, annotate and clean their data using custom scripts and ad-hoc methods before doing any analysis. However, research showed that all the steps preceding the analysis can introduce errors and misconceptions, especially since social scientists are often not trained in Computer Science and Data Science [?, ?]. Social scientists usually visualize their network using SNA tools like Gephi [4], Pajek [?], and NodeXL [?] which encompass basic interactions, node-link visualization, SNA measure computations, and clustering algorithms. Once they visualize their data, they typically notice errors and inconsistencies in the data, such as duplication of the same entities, merging of different entities, or geolocation errors. They, therefore, have to go back and forth between the visualization software and the cleaning process which can be tedious, especially since it can be complicated to trace back the entities of the data model back to the original documents for correction. VA tools which encompass the whole process of social historians should therefore be beneficial for the flow of their work and could help detect and correct errors or analysis plans way before the final analysis.

Moreover, social scientists have been frustrated by current data mining tools as the results they get from tasks such as clustering and classification come from black-box algorithms and may mismatch their vision of the data. They thus often feel they do not have control over their analysis [44]. One of the reasons is that automatic results can be hard to interpret, especially in disciplines such as History and Sociology, where users often have little knowledge of computer science.

1.4 . Contribution and research statement

As seen in section 1.3, most VA tools for HSNA are focused on the final analysis of a polished network. Instead of only providing analysis capabilities, We think VA could help social historians in their entire workflow, from data acquisition to analysis, instead of focusing only on the final steps. Concerning the analysis step, VA tools should satisfy the constraints of social historians, i.e. being able to

answer complex questions with effective visual representations, letting the user in control of the analysis loop and algorithmic results, and satisfying traceability to the original sources. The goal of this thesis is to give answers to the high-level question "How can VA help social historians in their entire HSNA process ?". To answer this question, we first characterize the HSNA process from start to finish, with the goal of identifying pitfalls and problems that regularly arise and characterizing social historians' needs. From this, we give answers and directions to three issues that frequently arise in social scientists' workflows :

Q1 : How to model historical documents into an analyzable network with the right balance between expressiveness and simplicity ?

Q2 : What representations and interactions are necessary for social historians to answer complex historical questions while being easy enough to navigate ?

Q3 : How to design VA tools and interactions that leverage algorithmic power but let historians in control of their analysis without introducing bias ?

In chapter 3, we start by describing the HSNA workflow and identify recurring pitfalls we encountered in our collaborations with historians and give answers to **Q1**. In the following chapter 4, we give answers to **Q2** by providing a VA interface to explore bipartite multivariate dynamic networks, with queries and comparison interactions with the aim of letting historians answer their questions. Finally, in chapter 5, we propose PK-Clustering, a clustering technique for social scientists based on their prior knowledge, algorithmic consensus, and traceability of results, as a concrete example of a technique giving answers to **Q3**.

2 - Historical Network Analysis and Visualization

Social historians rely on textual historical documents to draw socio-economic conclusions about the past. They read and analyze the documents they can find from a period and subject of interest, and make their conclusions after analyzing them and cross-referencing the information they found. Several methods have been developed in History to extract and analyze the information contained in the documents in a rigorous way, such as qualitative analysis, quantitative methods, or Historical Social Network Analysis (HSNA). HSNA is a method coming from Sociology consisting in modeling the relational information mentioned in the documents—such as family, business, or friendship ties—in a network, to be able to characterize and explain social behaviors through the description of the network's structure. HSNA is directly inspired by Social Network Analysis (SNA), which is a well-known method in sociology where a lot of methods and protocols had already been proposed when historians started to use similar approaches. Historians appropriated this method and adjusted it to the historian framework which differs from sociology by its relation to the sources (historians are limited by the documents they have, and all their work goes through them) and its focus on the temporality of their studies. 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 hypotheses once they successfully constructed their network. As the network 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. In this chapter, we therefore present a general overview of the fields of SNA (section 2.1), HSNA (section 2.2), and Social Network Visualization (section 2.3).

2.1 . 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 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 using networks. SNA is now a well-praised methodology in sociology, which has also been extended to History to study relational aspects of societies

and institutions of the past.

2.1.1 . Sociometry to SNA

One of Sociology's main goals is to study social relationships between individuals and find recurrent patterns and structures allowing to explain the behaviors of people and groups. Traditional methods try to explain social phenomena using classical social classifications such as age, social status, profession, and sex. For example, the socio-economic position of people living in a small city could be explained by their age, demographics, and family status which are traditional social categories. However, some criticism emerged that this type of division is often partially biased and comes from predefined categories which are not always grounded in reality. Sociometry is considered one of the bases of SNA and had the goal of redefining social categories through the lens of real social interactions and ties between persons, that sociologists wanted to observe in real conditions. It is in the 1930s that Moreno started to develop this new method by trying to depict real social interactions as a way to understand how groups and organizations were functioning [51]. He developed sociograms as a way to visually show friendships between people with the help of circles representing persons and lines modeling friendships. Figure 2.1 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 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 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.

2.1.2 . Structuralism and Ego Studies

Lots of sociological studies used SNA concepts after it has been formalized. However, there were 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 thought emerged with different objectives and methods : the structuralists and the school of Manchester [?, 19, 46].

The structuralists are interested in observing the relational structures and patterns forming a network, to make parallels between them and the social behaviors of actors in real life [?]. They think the positions of the persons in the network and the relational patterns they are part of reflect well the social activities and behavior in real life. Studying those would thus allow them to make interesting sociological

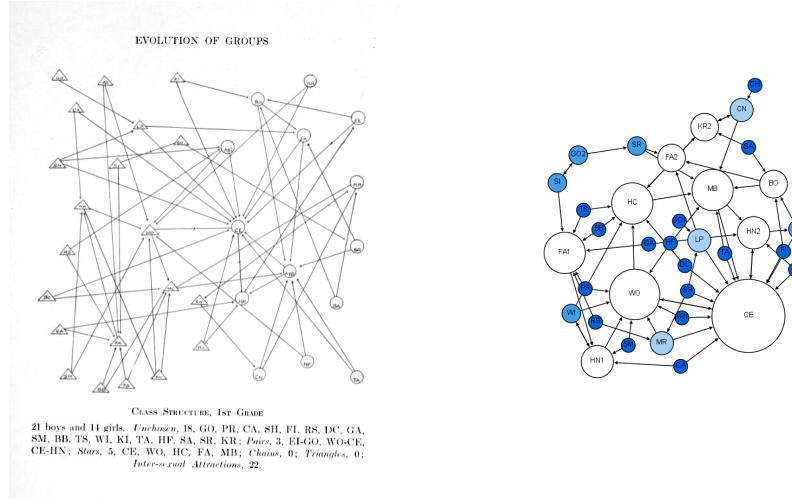


Figure 2.1 – Moreno’s original sociogram of a class of first grades from [50] (left). The diagram shows 21 boys (triangles) and 14 girls (circles). The same sociogram plot using modern practices generated from Gephi from [27]. The color encodes the number of connections incoming.

conclusions. Accordingly, sociologists in this school usually study organizations and specific groups—such as institutions, companies, families, etc.—and want to explain their functioning through the description of the internal shapes and structures of the networks. Thus, they try to construct networks that exhaustively model all the interactions between the actors constituting the groups, as missing links would misrepresent the reality of interactions.

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 [47] where he showed that the integration of urban mores and customs were directly correlated to the persons’ relationships networks in the city. Xhosa¹ 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

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

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. 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.

2.1.3 . Methods and tools

Graph theorists and network scientists developed a myriad of measures and algorithms that social scientists appropriated to describe and characterize social phenomena. When constructing networks, the first thing sociologists do is often to identify the main actors of the network and explain why these actors are the most central, for example by linking it to their profession or social status. Computing the degree—which is the number of connections for a node—distribution is the main straightforward way of doing it, but other more complex measures like centrality have also been developed. Lots of types of centrality have been proposed, based on different criteria, as there are several ways of defining the more important actors. Some centrality measures highlight actors with the highest number of connections while others highlight people bridging different groups with low interactions. More generally sociologists 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 into low-level relationships between people. This reflects on Simmel's formal sociology, where he already referred to dyads and triads as a primal form of sociability [?]. More recently, graphlet analysis extended this concept to every pattern of N-entities. Graphlet analysis aims at enumerating every small structure of N nodes composing a network, to understand how people interact at a low level. Graphlets counting shows that graphlets are not found in a uniform distribution in social networks, thus revealing that social networks usually do not have the same structure that random networks. This is a fact well known in SNA. Precisely, entities in real-world networks tend to agglomerate into groups (also called clusters) where entities in the same groups interact more between them than with entities from other groups. From a sociology perspective, it means that people tend to interact and socialize in groups and interact more rarely with other people from outside groups. These groups are often referred to as *communities*, and a lot of algorithms have been proposed to find these automatically.

2.2 . 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 in the 1960s, with the rise of social history, by extracting information from historical textual documents and studying them with statistical methods. When seeing the potential of SNA concepts for historical purposes, historians started to extract the relational information contained in documents to study historical social phenomena using the power of networks and methods already developed in SNA.

2.2.1 . History, Social History and Methodology

History is the science of retrieving and characterizing facts about the past. It emerged as a field with its own rules, conventions and journals in the 1880s from faculties of letters, to counterbalance previous history works which were judged as too "literary" [?]. History can be seen through two facets : one is societal, and serves creating a shared story for the country and a sense of unity to its citizens. Antoine Prost says that "it's through history than France thinks itself" (translated from french) [60]. The second facet of history constitute a methodology to describe the past in a rigorous and scientific way, with proofs. For this, historians rely on historical documents that they leverage to infer dated facts about the past (the temporal aspects of conclusions is always central to the historian work). The textual sources are thus at the core of the work of the historians, and having to cite historical documents and previous peers work to new claims is primordial to be considered as rigorous History work. However, even if those two aspects are well characterized (temporal aspect of the work and its relationships to sources), methodological and epistemological facets (how historians should read and analyze their sources, how to cite them, what to report/not report etc.) of History have not been studied and discuss for a long time, until the end of the 1980s. Some historians were interested in historiography [?], but none were going to philosophical and epistemological reflexions of the History discipline. For Lucien fèbvre, philosophising was even constituting a "capital crime" [?, 60].

alexis: maybe talk about the positivists and methodists At the start, history was mainly event-centered, and was focusing in characterizing central figures of the past like rulers and artists or shed light on events which shaped history like wars or political crisis. This narrative approach to history has been criticized for its open interpretation of historical documents, which can introduce bias from the authors [?].

In the 1930s, March Bloch and Lucien fèbvre detached from traditional history by creating the "Annales school" (Ecole des Annales) which tried to replace the human as a component of a broader sociological, political and economic system with influences between each other [?]. 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 got successful in a profession in crisis, by bringing a new lens of study on various societal subjects more grounded in the

real and with a better intelligibility. This school of thought can be seen as one of the biggest milestones for Social History, a branch of History which focuses on the socio-economical aspects of societies and their changes through time, rather than an event-centric view of History. For example, in his thesis, Ernest Labrousse—a well known figure of Social History—tries to describe and explain the economic crisis of France at the end of the “Ancien Régime”² through the evolution of the economic power of different social groups such as farmers, workers, property owners etc instead 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 historic facts through a complex and social aspect and with a strong focus on sources and traceability.

2.2.2 . Quantitative History

With the development of statistical methods and more precisely Computer Science, quantitative approaches of History emerged in the 1960s with the aim of analyzing quantitative data directly extracted from historical documents. Using such methods, historians were able to make conclusions based on statistical results on topics such as demography [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 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 Arriage went as far as criticizing the decrease of interest of historians 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 losing the trace of the original sources). It can also provide infrastructures and tools to study large historical databases which is more complicated to do by hand, as with the Venice Time Machine project [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.

2. The “Ancien Régime” is an historic period of France which starts from the beginning of the reign of the Bourbon house at 1589 until the Revolution in 1789.

2.2.3 . Historical Social Network Analysis

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 us to model those relational entities more thoroughly using network concepts, thus allowing us to make new observations that it was not possible to see without taking into account the relational aspects of these entities. Observing and describing the structure of the resulting networks allowed historians to make conclusions on sociological aspects of the past, similar to SNA. Since then, HSNA has been applied by historians to study multiple kinds of relationships, like kinship and political mobilization [45], administrative and economic patronage [53], etc. If these approaches fall under similar critics of quantitative history [41] like 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 2.2 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 HSNA methods 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].

2.2.4 . 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 of vertices representing the actors of interest (very often individuals mentioned in the documents), and $E \subseteq V^2$ a set of edges modeling the social ties between pairs of actors. This allows to have a simple network to visualize and analyze, but it does not always reflect the sociological complexity of information contained in the documents. 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

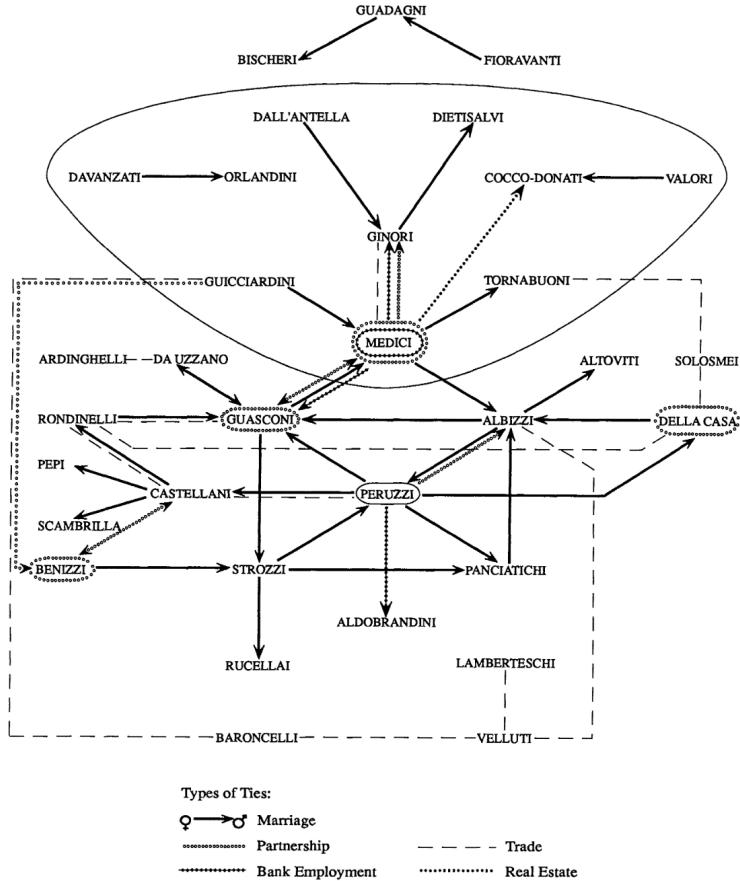


Figure 2.2 – Marriage, partnership, trading, banking, and real estate networks of the powerful families of Florence from [57]. We can see the central position in the network of the Medici Family.

where the relations link employees to organizations but not employees to employees or organizations to organizations. Many social situations or documents can be modeled in these terms (affiliation lists or co-authoring). Multivariate networks, i.e., graphs, where vertices and edges can be assigned multiple “properties” or “attributes”, are less used in SNA. These attributes are often considered secondary, the emphasis of SNA being on the topology, its features, measures, and evolution.

Historians, demographers, sociologists, and anthropologists have also been designing specific data models for their social networks, based on genealogy or more generally kinship [?]. For genealogy, the standard GEDCOM [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].

2.3 . Social Network Visualization

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

2.3.1 . Visualization

Data Visualization consists in graphically displaying data for 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's 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's 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 "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 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, 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 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 2.3.

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 Analy-**

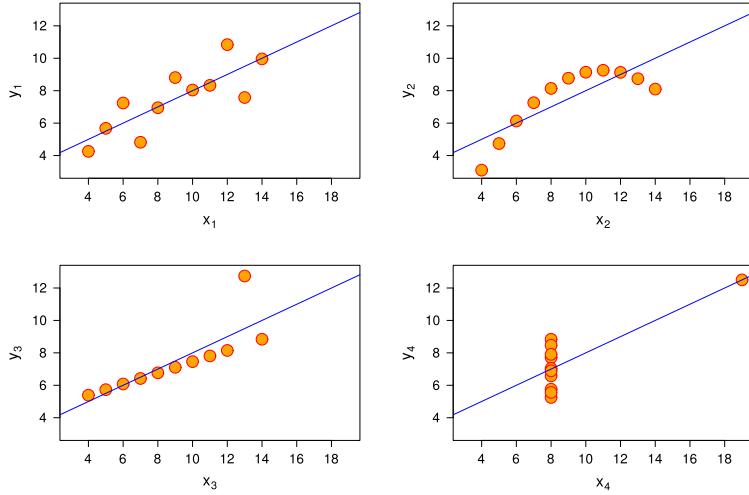


Figure 2.3 – Anscombe quartet. The four datasets have the same descriptive statistics (average, variance, correlation coefficient) but very different structures. Image from [3].

tics emerged later from Information Visualization by mixing data mining and more complex analysis process with traditional information visualization displays. 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.

2.3.2 . Social Network Visualization

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 [?] 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 [10, 40]. Figure 2.4 shows some of these metrics, synthesized by Kosara and al. [40]. In Figure 2.1 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 the most

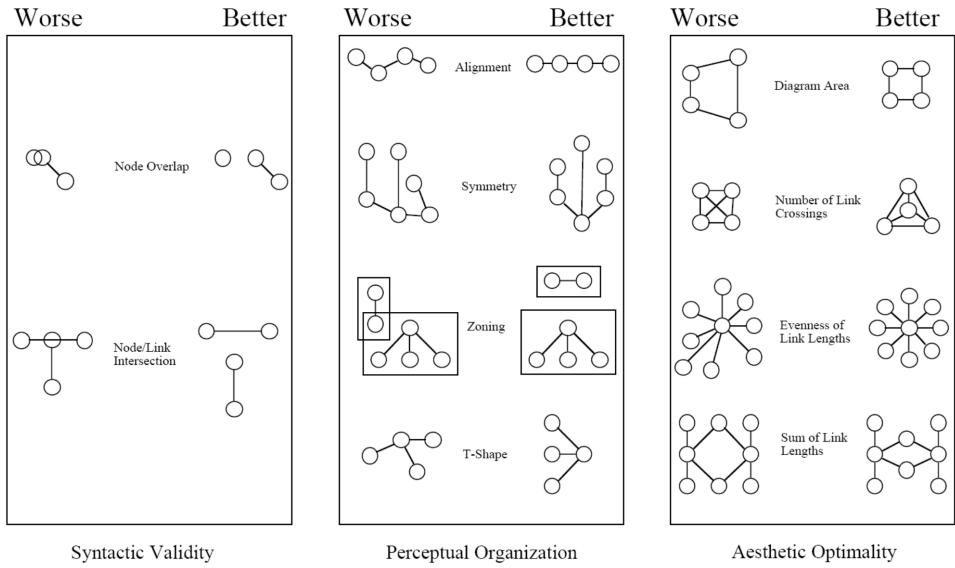


Figure 2.4 – Different criteria are proposed to enhance node-link diagram readability. Image from [40]

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. Lots 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 [49]. Still, Matrices have been shown to be better than node-link diagrams 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 2.5), geolocated social networks with the Vistorian [?], and multivariate networks with Juniper [55]. However, these new network representations take time to be adopted by social scientists who rarely use those.

2.3.3 . Social Network Visual Analytics

Social network visualization has mostly been used for confirmatory and communication purposes from its beginning. Social scientists often had hypotheses 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 into the data and potentially generate new hypotheses. This process has been character-

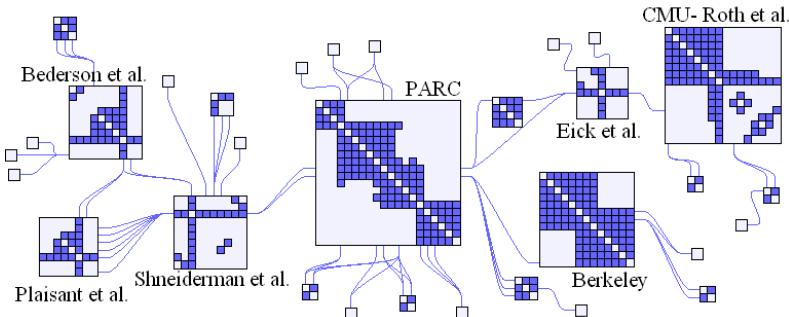


Figure 2.5 – NodeTrix system showing a scientific collaboration social network with clusters. Each cluster is represented as a matrix, Image from [33].

rized by Tukey in 1960 as *Exploratory Data Analysis (EDA)*. Exploration is mostly possible thanks to interaction, which allows changing 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 hypotheses.

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 with graphical displays, letting users apply data mining algorithms directly in the exploratory loop. This coupling of visualization and data mining has been defined as Visual Analytics (VA) and is still undergoing lots of research. Keim and al. define it as "a combination of automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data" [38]. Figure 2.6 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 generates and explores 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 hypotheses. The most used social network VA tools are Gephi [4], Pajek [?], and NodeXL [?]. Figure 2.7 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

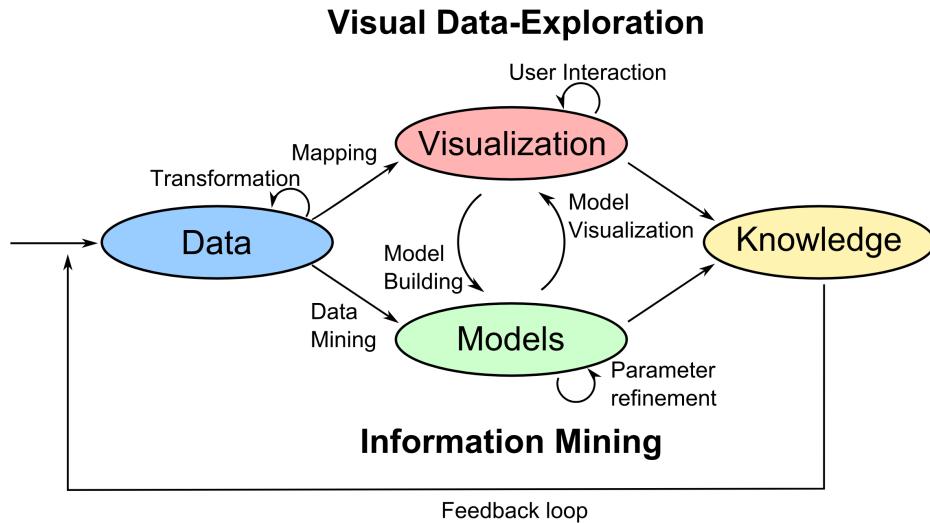


Figure 2.6 – Abstraction of the VA process. It is characterized by continuous interactions between the data, visualizations, models, and knowledge. Image from [38].

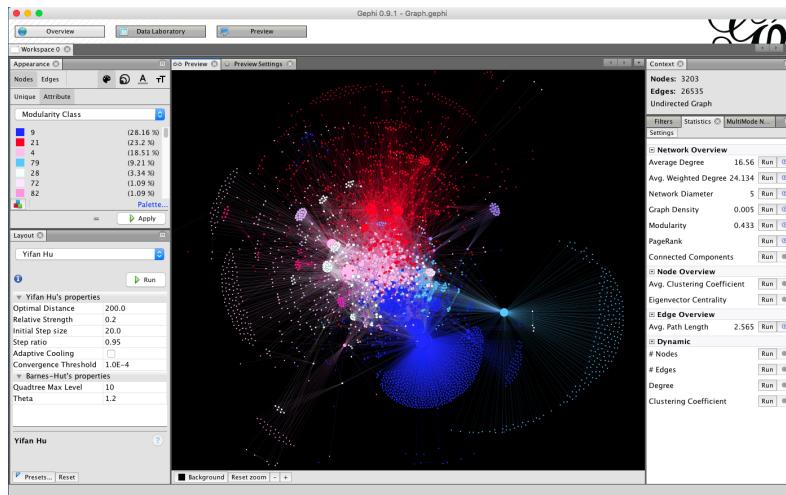


Figure 2.7 – Gephi [4] interface. The network is represented with a node-link diagram. Users can interact on the visualization and encode node and links visual attribute (color, size etc.) with network measures computed directly in the interface, such as the node degree, or clustering results.

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 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 go back and forth between their analysis process inside the VA tool they are using, and their original sources and annotation/modeling process, to correct errors or modify annotations. Interestingly, the chosen network model plays a big role in 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 interfaces enforce simple network models, making this retroactive process harder. Some interfaces still incorporate data models encapsulating document representations, such as Jigsaw [?] 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 to help them to do easier back and forth between their analysis and the annotation, network modeling, and cleaning steps, as they play a big role in the historian workflow.

3 - HSNA Process and Network Modeling

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

3.1 . Context

Tools for social network visualization tend to ignore the context in which the networks are produced, where they come from, and the workflow that led from their origin (e.g., documents, polls, interviews, web scraping) to their network form. Yet, practitioners of social history need to generate many networks from the same documents/sources to visualize and analyze them. In this chapter, after describing and characterizing the workflow of Historical Social Network Analysis [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*. These principles emerged from joint experiences as historians and computer scientists while collaborating on multiple projects.

Social historians' goal is to characterize socio-economic phenomena and their dynamics in a restricted period and place of interest and to see how individual people of that time lived through those changes. For this, they rely on historical documents such as conversational letters, censuses, and marriage acts. They usually extract qualitative and quantitative information from an identified corpus of documents, to then make conclusions on interesting socio-economic topics such as migrations, business dynamics, education, and kinship. For doing this, historians can apply HSNA methods, by modeling the social relationships between a set

of entities—usually individuals—into a network. Historians therefore collect documents, annotate them, construct a network from the annotations that they finally analyze and visualize to validate or find new hypotheses. Unfortunately, the process is often linear, and it is common that, when visualizing their network, historians spot errors and inconsistencies in the annotations that they could have fixed if the process was iterative.

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

3.2 . Related Work

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

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

With similar aims, Glaser and Strauss developed the “Grounded Theory” [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, correlations, statistical distributions, statistical testing, time series etc. [35, 42]. Similarly, the use of network science for historical aims increased in recent years, and a lot of resources exist on how to use network methods and measures for historical research [39, 41].

However, little work has been done on describing and formalizing the process before the analysis part for a quantitative and network research workflow. Indeed, if it is central to know how to manipulate statistical and network concepts and methods when following this kind of methodology, it is as important if not even more to follow a correct and rigorous workflow to generate the data we plan to analyze beforehand. The process to generate a clean quantitative or network dataset from historical sources is difficult and requires several data acquisition, annotation, and cleaning steps. Social analysts are not always trained on how to do these steps effectively, which can lead to errors, inconsistencies, and mismatches between the chosen data models and the historical questions [2]. Karila-Cohen and al. provide some advice on how to annotate historical documents with the aim of using quantitative methods [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 choose a network model which fits their analytical goals.

3.3 . Historical Social Network Analysis Workflow

From the literature and our own projects of HSNA we conducted during the last years in collaborations with historians, we propose an HSNA workflow divided into 5 steps : *textual sources acquisition*, *digitization*, *annotation*, *network creation*,

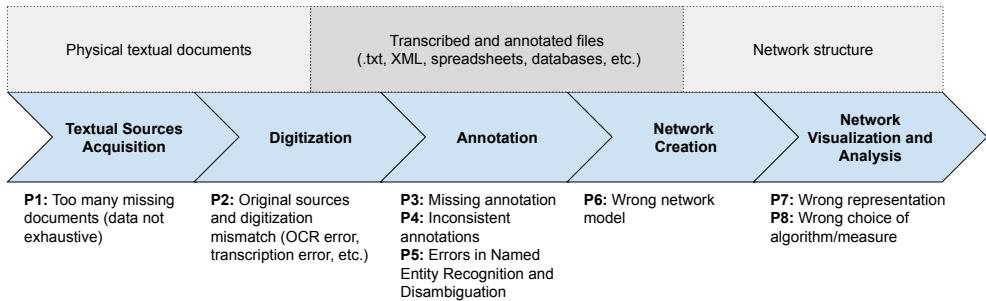


Figure 3.1 – HSNA workflow is split into five steps : textual sources acquisition, digitization, annotation, network creation, network visualization and analysis. We list potential pitfalls for each step.

and finally *visualization and analysis*. The workflow is presented in Figure 3.1 along with potential and recurrent pitfalls.

3.3.1 . Textual Sources Acquisition

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

3.3.2 . Digitization

Digitization consists in converting the sources into a digital format. This step can be skipped for the most recent periods where many documents have been produced digitally or can be scanned and well digitized through optical character recognition (OCR), allowing to tremendously ease the storage, indexation, and annotation of the documents. However, before mid 20th century, most historical primary sources are stored in archives in paper format and need human work to be digitized. **Mismatches between the original documents and the transcription can occur for old and recent documents (P2)**. However, if OCR tools are more and more efficient in English and highly used languages, historians can work with old documents written in old or extinguished languages and with atypical writings (e.g., Fraktur handwriting and typefaces for German in the early 20th century). Therefore, OCR tools are often unusable in social history and digitization remains

an expensive and sometimes highly skilled process.

3.3.3 . Annotation

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

3.3.4 . Network Creation

Historians construct a network from the annotations of the documents. Usually, all persons mentioned are annotated and will be transformed into network nodes

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

3.3.5 . Network Analysis and Visualization

Once historians have constructed a satisfactory network, they start exploring and analyzing it with visualization and quantitative methods. The final goal of HSNA is to find interesting patterns and link them to social concepts to gain high-level socio-historical insights [?, 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 structures 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 socio-logical meanings. If we take as an example betweenness centrality which measures the number of times a node appears in the shortest path of every pair of existing nodes, individuals with high values usually highlight positions of power as they communicate with different groups. However, it can also be interpreted as a position of vulnerability in other contexts such as during periods of wars and repressions, as in the study of Polish social movements in the 20th century by Osa [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.

3.4 . Network modeling and analysis

Historians typically construct one or several networks from their annotated documents that they will visualize and analyze to validate or find new hypotheses. As the processing steps of the workflow are often not transparent (digitization, annotation, network modeling), it can be difficult for the reader of an HSNA study to understand how the network has been constructed, what it represents, and to trace back the network entities to the original sources [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.

3.4.1 . Network Models

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

- **Simple Networks** [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 modeling more complex social relations where people can have various social ties e.g. as

parents, friends, and business relationships. However very often several possible representations for the same data exist as projections are often applied to the original documents to get this type of model. One of the main drawbacks of this model is that it creates parallel edges that are hard to visualize.

- **Bipartite (also called 2-mode) Networks [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 analyses in HSNA encode the *roles* of the persons in the documents as link types [?]. This network model is more aligned with the original sources and allows following an analysis through the original documents themselves and not through concepts. For example, the GEDCOM format introduces the concept of “family” that ties together a husband, spouse, and children with different link types. However, the concept of family can have different meanings across time and cultures, meaning that GEDCOM adds a conceptual layer instead of grounding the network to concrete traceable documents and events (e.g., no marriage but birth certificates).
- **Multilayer Networks [48]** : in these networks, each node (vertex) is associated with a *layer l* and becomes a pair (v, l) , allowing to connect vertices inside a layer or between layers. These advanced networks have received attention from sociologists [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) [?]** : 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.

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

Currently, most digital historical projects use unipartite networks (simple, co-occurrence, and multiplex) that are simple and allow answering specific questions, but they do not capture all the complexity of the documents, and social scientists

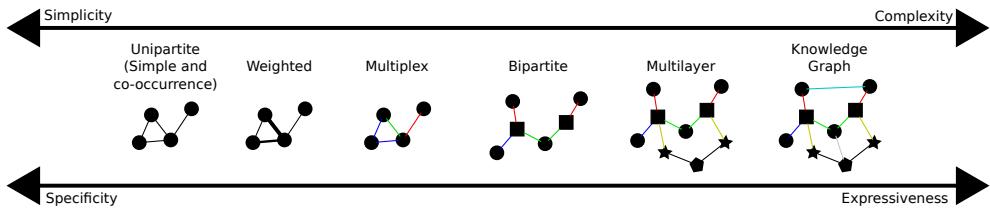


Figure 3.2 – Schematic representations of Different network models used for analyzing historical documents, ordered by complexity and expressiveness

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

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

3.4.2 . Bipartite Multivariate Dynamic Social Network

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

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

Links and Roles : A link models the mention of a person in a document. **Each link has a type corresponding to the role of the person in the document**. For a marriage act, the roles include *wife*, *husband*, *witness*. This is a key aspect of our model since it clarifies the relationship between the persons within an event. In contrast, Jigsaw [?] does not consider the roles.

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

(activity, residence, etc.).

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

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

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

3.4.3 . Examples

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

1. Analysis of the social dynamics from **construction contracts in Italy in the 18th century** [?, 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 guarantees, and *Approvers*, who vouch for the guarantors. Documents contain information about the building site, the type and materials of constructions, and the origin of the people.
2. Analysis of migrations from the **genealogy of a french family between the 17th–20th centuries** [unpublished work]. The corpus is made of family trees referring to several document/event types : birth and death certificates, marriage acts, military records, and census reports. The roles are different for each event type and consist of *children*, *father*, *mother* for the birth events, *deceased* for the death event, *spouse* and *witnesses* for the marriages, and *family members* for the census events.
3. Analysis of migrations from Spain to Argentina through the **marriage acts at Buenos Aires in the 17–19th centuries** [?, 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)** [?].

The corpus is made of administrative forms that mention persons requesting to migrate, along with the persons they want to join, and the administrative persons of the ministry in charge of the forms. The family members of the aspiring migrants are also mentioned in the forms, with their respective dates of birth.

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

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

Moreover, projections add ambiguity in retrospect of the original documents, as it becomes impossible to trace back one link to one specific document, as the same link could potentially refer to several ones [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 information like a date that we model as an attribute, we can store it as a document node attribute using a bipartite model. However, when projecting the network this information appears in the links as many times as there are persons mentioned in the document minus one and often more. For example, in the example #1 in Table 3.1 the time is stored in $\sum_{i=1}^4 i = 10$ links in the co-occurrence model and in 9 links in the multiplex unipartite model while it is only stored once as a document node attribute in the bipartite model.

3.5 . Applications

Several tools have been designed for visualizing dynamic bipartite networks that can also be considered dynamic hypergraphs [58, 70], but few incorporate attributes.

Original Document	Co-occurrence	Unipartite representation	Bipartite
<p>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).</p> <p>Husband Wife Witness</p>	<pre> graph TD T1((T1)) --- H((H)) T1 --- W((W)) H --- W H --- T2((T2)) W --- T2 1659((1659)) --- H 1659 --- W 1659 --- T2 1659 --- T1 </pre>	<pre> graph TD H((H)) --> W((W)) H --> T1((T1)) H --> T2((T2)) W --> T1 W --> T2 T1 --> T2 1659((1659)) --> H 1659 --> W 1659 --> T1 1659 --> T2 </pre>	<pre> graph TD H((H)) --- M[1659] W((W)) --- M T1((T1)) --- M T2((T2)) --- M </pre>
<p>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</p>	<pre> graph TD G((G)) --- A1((A1)) G --- A2((A2)) G --- A3((A3)) A1 --- A2 A1 --- A3 A2 --- A3 A1 --- Ap((Ap)) A2 --- Ap A3 --- Ap 1712((1712)) --- G 1712 --- A1 1712 --- A2 1712 --- A3 1712 --- Ap </pre>	<pre> graph TD G((G)) --> A1((A1)) G --> A2((A2)) G --> A3((A3)) A1 --> A2 A1 --> A3 A2 --> A3 A1 --> Ap((Ap)) A2 --> Ap A3 --> Ap 1712((1712)) --> G 1712 --> A1 1712 --> A2 1712 --> A3 1712 --> Ap </pre>	<pre> graph TD G((G)) --- M[1761] A1((A1)) --- M A2((A2)) --- M A3((A3)) --- M Ap((Ap)) --- M </pre>
<p>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, âgée de vingt-six ans , demeurant au village de Grudet, dite commune de Saint-Symphorien. [...] Father Mother Child</p>	<pre> graph TD F((F)) --- M((M)) F --- C((C)) M --- C 1901((1901)) --- F 1901 --- M </pre>	<pre> graph TD F((F)) --> M((M)) F --> C((C)) M --> C 1901((1901)) --> F 1901 --> M </pre>	<pre> graph TD F((F)) --- M[1901] M --- C((C)) </pre>

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

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 [?], as it would allow tracing the entities of the network back to the original documents more easily. Mechanisms to clean/modify the annotations and reflects on the network modeling process directly in the analytical environment could also ease the social scientists' workflow loop. It would allow them to directly clean errors and inconsistencies in the annotations and propagate them in the visual analysis workflow. For example, the Vistorian [?] now lets users modify and clean their data in a table format if they see errors or inconsistencies.

3.6 . Discussion

Most tools for social network visualization focus solely on the visualization and analysis steps, without considering the whole historical data analysis process, preventing researchers from going back to the original source, and supporting the social analyst in the annotation and modeling steps. We think visual analytics tools helping social scientists annotate and model their data with *reality*, *traceability*, and *simplicity* principles in mind are essential to conducting socio-historical inquiries with limited friction, realistic training, and scientific transparency. Concerning the network modeling step, bipartite multivariate dynamic networks model well the majority of structured historical documents such as marriage acts, birth certificates, and business contracts as these documents refer to specific events (birth, marriage, transaction, etc). The document nodes, therefore, represent both the textual documents and the specific events. This dual representation works well for semi-structured documents but could be more limiting for other more literary documents. Moreover, structured documents can also provide information about other relationships not directly linked to the main event. For example, marriage acts sometimes refer to the place and date of birth of the spouses with the names of the parents. This information relates to the birth of the spouses and not the marriage specifically. In that case, social historians can either ignore this type of information in the annotation process or encode it with specific roles (*husband's father* and *wife's father* for example), thus turning the network into a model of the documents only, and not events. We show what would look like the resulting networks Figure 3.3 for the two cases where marriage acts mention birth information and the case where only marriage-related information is present in the document.

3.7 . Conclusion

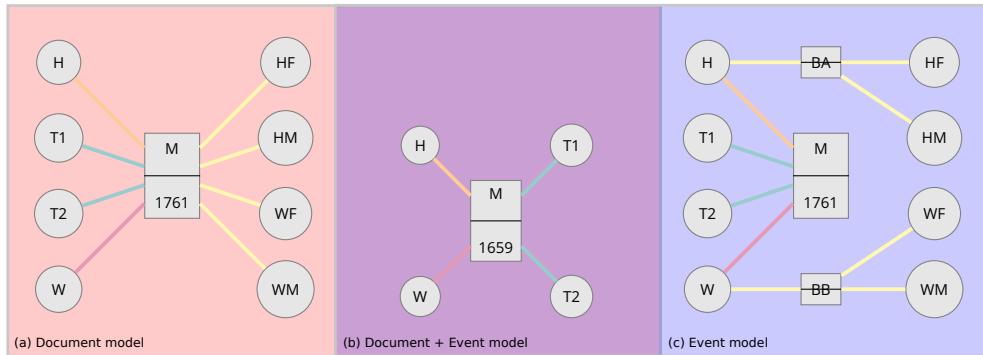


Figure 3.3 – bipartite multivariate dynamic network modeling for two cases of marriage acts of example #3. Some marriage acts mention the parents of the spouses, which is a relationship different than the marriage in itself. This case can be modeled using a document model (a) or an event model (c) by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

HSNA is a complex process that starts by collecting historical documents and ends with elaborating high-level sociological conclusions. Historians support their conclusions by modeling individuals' social relationships extracted from the documents and analyzing the resulting networks. We tried to shed light on this process by dividing it into 5 steps and describing recurrent pitfalls we encountered in our projects and collaborations. More importantly, we think this process should be done following the principles of *reality*, *traceability*, and *simplicity*, to avoid biasing the analysis, allowing to go back to the original source at any point of the workflow, and using models and methods simple and powerful enough for social scientists. Visual analytics software designed for HSNA should consider those principles to provide tools allowing to follow non-biased and reproducible analysis starting from the raw documents while supporting historians in going back and forth more easily between the annotation and analysis/visualization steps. We discussed the network modeling process in depth and claim that bipartite multivariate dynamic networks satisfies those three core principles, letting historians both wrangle their data and characterize sociological phenomena using a common model and visual representation. Therefore, using this model VA interfaces could help social scientists manage and analyze their data starting at the data acquisition and annotations steps instead on focusing on the analysis only, while providing efficient representations of the data for analysis and exploration. We explore what could be such VA interfaces in the two next chapters.

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

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

4.1 . Context

Social historians and sociologists aim at retrieving and studying facts about a specific region and period of time that they focus on. Their work essentially relies on documents—such as marriage acts, census records, surveys, and business contracts—to gather information about the life of important actors that they explore in-depth, or to draw conclusions on social aspects of groups in the society of that period and place. Instead of drawing conclusions from their gathered knowledge and interpretations of the documents, a more systematic approach consists in constructing a social network from the documents and following a Social Network Analysis (SNA) approach [72]. For this, they need to encode their documents to extract the persons and any other useful information in the text and transfer it into a structured file or a database. Social scientists can then explore, validate, or refute their hypotheses by observing and analyzing the network structure and the connectivity patterns between the entities of the resulting network. They also want to visually explore their data to generate new insights and hypotheses.

Currently, social scientists often model their datasets as simple networks where the nodes are the persons mentioned in the documents (see chapter 3). Usually, Two persons are then connected together in the network when they appear in shared documents. This representation is easy to visualize and analyze but simplifies and distorts the information by hiding the documents that witness the relationships between the persons. Thus, another approach consists in modeling the data as bipartite networks, where both the documents and the persons are represented as nodes and are connected together when a document mentions a given person [?, ?, ?].

In addition, historical documents include time and geospatial information cor-

responding to the date and location of the events they refer to. Documents often mention additional information on the persons, such as their sex, profession, and date of birth. These are often essential to understanding underlying social phenomena, as time, space, and social status play an important role in social dynamics. For these reasons, historical sources and the underlying social phenomena they refer to can be modeled well by *bipartite* with *roles*, *multivariate dynamic* networks. *Bipartite* means that both persons and documents (or events, that are often witnessed by physical documents) are modeled as typed nodes. *Multivariate* means that the nodes and links can carry additional attributes. *Dynamic* means that time is a mandatory attribute of documents. Furthermore, a link created between a person's node and a document's node (when the person is mentioned in the document), has an associated link type that models the *role* of the person in the document/event. Additionally, documents can optionally carry a geographical location. This model unifies several social network models and allows to model the historical sources with any transformation, simplification, or loss of information [9].

Several sophisticated tools exist to explore and analyze rich social networks. However, the majority of them either enforce too simplistic network models, such as Gephi [4] and NodeXL [?] or do not enforce any data model and lead to very complicated interfaces which are complicated to navigate for users like historians. Moreover, the majority of social network visual analytics tools provide limited interactions to query and explore richly encoded data, and historians often reach simple conclusions.

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

we thus focus on three high-level tasks : exploration, queries, and comparison of this type of network. Users can explore the data using two layouts : a node-link bipartite view showing the sociological structure of the network, and a map layout based on the geolocation of documents. We designed and implemented a new visual graph query system that allows us to build both topological and attribute constraints, based respectively on a node-link interactive representation, and dynamic widgets. For this, we rely on the Neo4j graph database [?] and its language *Cypher*. Most visualization systems offer dynamic queries to hide the complexity of query languages. However, using a rich data model, some queries are much easier to refine using scripting than dynamic queries. We implemented dynamic queries that also show the translated Cypher queries, and inversely, can translate textual queries into visual queries. With that interface, social scientists can start building

their queries with simple widgets and, if needed, complement them by editing the query, alone or with the help of power users. On top of that, they can easily copy and paste the textual query to share the current state of the query and associated results with someone else or to start an analysis session from a previous result. ComBiNet also implements subgraph comparison techniques, allowing the comparison of networks, network-related measures, and attribute distributions between the entities returned by the queries. We validate the query and comparison system with a formative usability study and we demonstrate ComBiNet can be used to answer sociological questions by describing in depth several real-world use cases.

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

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

4.2 . Related Work

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

4.2.1 . Graphlet Analysis

One of the inspirations for this project came after participating in the 2020 VAST challenge¹ where we used graphlets to measure similarity between several networks [?].

Graphlets are small connected induced, non-isomorphic subgraphs composing any network. In an induced subgraph, two vertices linked in the original graph remain linked in the subgraph. For instance, if the original graph is a triangle we

1. This is a challenge organized in the context of the IEEE Visual Analytics Science and Technology (VAST) conference. The challenge consisted of a series of analytical questions united under an overarching cyber threat scenario. We participated in the Mini-Challenge 1 which asked participants to identify a group of people that accidentally caused an internet outage. To identify this group, we were given a network profile and a large multi-variate social network to search in.

can only induce the simple edge or triangle subgraph (graphlet). The path of length 2 has all vertices of the original graph but misses an edge and is, therefore, not a possible graphlet. They were first introduced by Milo et al. [?] to explore the structural differences between biological networks, but they are now used in several disciplines involving networks such as sociology.

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

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

4.2.2 . Visual Graph Querying

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

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

4.2.3 . Visual Graph Comparison

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

Andrews et al. [?] describe a technique to compare several graphs, using a

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

4.2.4 . Provenance

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

4.3 . Task Analysis and Design Process

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

4.3.1 . Use Cases

We elaborated this interface from the collaborations with historians we described in subsection 3.4.3. These collaborations involved regular meetings and multiple interviews over two years. All these datasets are textual corpora constituted of historical documents mentioning people with complex relationships. They are well modeled by bipartite multivariate dynamic network. We give more details about

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

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

the datasets of these collaborations in this section and we also list our collaborators' main questions and the graph queries extracting the information to start answering them. The full answers involve visualizations of the query results and attribute summaries that we describe in the next section. We list the most important questions our collaborators shared with us on their respective datasets. We categorized those according to four dimensions : global (G)/local (L) (do they want to categorize a group of nodes or retrieve specific persons/documents), if the question can be answered using the topology (T), and/or the attributes (A), and finally if a comparison (C) using several filters is needed or not (N).

1. Analysis of the social dynamics from **construction contracts in Italy in the 18th century (141 documents, 272 persons)** [?]. The corpus is made of contracts (manuscript documents) for different types of constructions in the Piedmont area in Italy. People are mentioned in three different roles : *Associates*, who participate in the construction ; *Guarantors*, who bring financial guarantees ; and *Approvers*, who vouch for the guarantors. Along with the time and location of the construction site, documents have a construction type (military, religious, and civil), work type (big work, small work, reparation, transportation, etc.), and material (wood, stone, metal). People also have an origin attribute (the place they come from), manually extracted from the original documents.

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

Query 1.1 Request all approvers occurrences

Query 1.2 Request all associates and guarantors occurrences

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Query 9.1 Select persons with a geolocated birth and death certificate

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.3.2 . Tasks Analysis

Most of the questions we collected from our collaborators could be answered by isolating a subgroup of entities and analyzing them in the context of the whole network, or by comparing two subgraphs, in terms of their entities, structure, and attribute distributions. From discussions with our collaborators and the analysis of

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

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

their questions on their data, we elaborated a list of requirements for the visual interface, split into three main parts : 1) Exploration of the data, 2) Queries, and 3) Comparisons. The elaboration of the tasks was an iterative process, as we showed the interface to our collaborators several times in the development phase to get feedback. The tasks are described here and summarized in Table 4.1 :

1. **Exploration of bipartite multivariate dynamic network.** The visual interface must allow exploration of this specific type of network, using every aspect of the data, i.e. its topology (T1.1), node attributes (T1.2), roles (T1.3), geolocation of the documents/events (T1.4) and time (T1.5). Common interactions such as selection and zooming are also needed for the exploration.
2. **Applying filters.** To answer their questions, users need to be able to apply filters to the data, to isolate specific groups of entities having specific behaviors or characteristics. To answer the diversity of questions, they should be able to put constraints on every aspect of the data, i.e. the topology, the roles (T2.1), and the attributes (including time and geolocation) (T2.2). Access to provenance information can also help them in their query construction, by going to previous states and exploring different paths more easily (T2.3). Once they are satisfied with their query, they want to explore the results, usually in the context of the whole network (T2.4).
3. **Comparison of several subgraphs.** Users should be able to compare several subgraphs isolated after applying filters, to see the similarities and differences between groups of entities of interest. The system should be able to easily see the common and shared entities of the two subgraphs (T3.1), their respective place in the network, their structural differences (T3.2), and their different attribute distributions (T3.3).

4.4 . The ComBiNet System

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

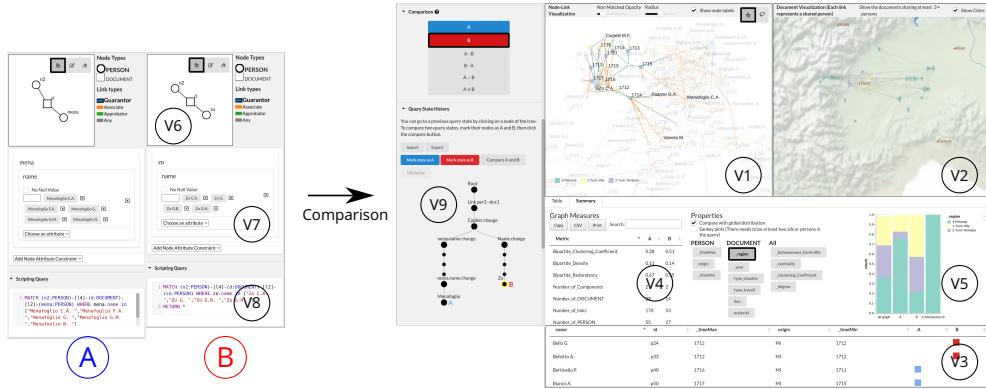


Figure 4.1 – The ComBiNet system used to compare two subgroups of a social network of contracts from [?], extracted with dynamic visual queries. (A) and (B) show the two visual queries created by the user in the query panel using an interactive node-link diagram editor (V6), dynamic query widgets (V7),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

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

4.4.1 . Visualizations

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

V1 : Bipartite Node-Link Diagram The bipartite node-link visualization panel shows the network using the DrL force layout from igraph [?] with overlap removal using D3 [?]. Node-link representations are very common in social sciences [?, ?, 4] and were a specific request from our collaborators. In the context of our bipartite model, the persons are represented as circles and the documents/events as squares, while the roles are encoded as link colors. A link models the mention of a person in a document. This view provides an overview of the data by showing the structure of the network (T1.1) and the roles of the persons in their different documents (T1.2). Attribute values can be overlayed on the nodes using colors when users

select an attribute. It allows detecting patterns relative to attributes, in the context of the topology of the network (T1.2, T1.4, T1.5). For example, Figure 4.2 shows the construction dataset of #1 where the user selected the *year* attribute, coloring the documents nodes with their year in the node-link diagram (left). The view also provides pan & zoom and selection interactions for effective navigation.

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

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

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

V3 : Entities Tables All the persons and the documents of the loaded dataset are listed in two separate tables, showing the attributes of the entities. This way, users can order the entities according to any attribute they want (T1.2). The tables are linked to the visualizations, meaning that selecting a row highlights the respective entity in the visualizations, and vice-versa. Tables in social network visualization systems have been proven to be efficient and useful for social scientists when exploring their data [?]. It allows them to link the visualization to the network entities more easily, and dive deeper into one entity's attribute values after selecting it in the network. It also makes ranking entities according to various criteria easier and more straightforward.

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

V5 : Attributes View All the attributes in the network are shown as buttons in the bottom right of the interface, sorted by their associated node type (person,

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

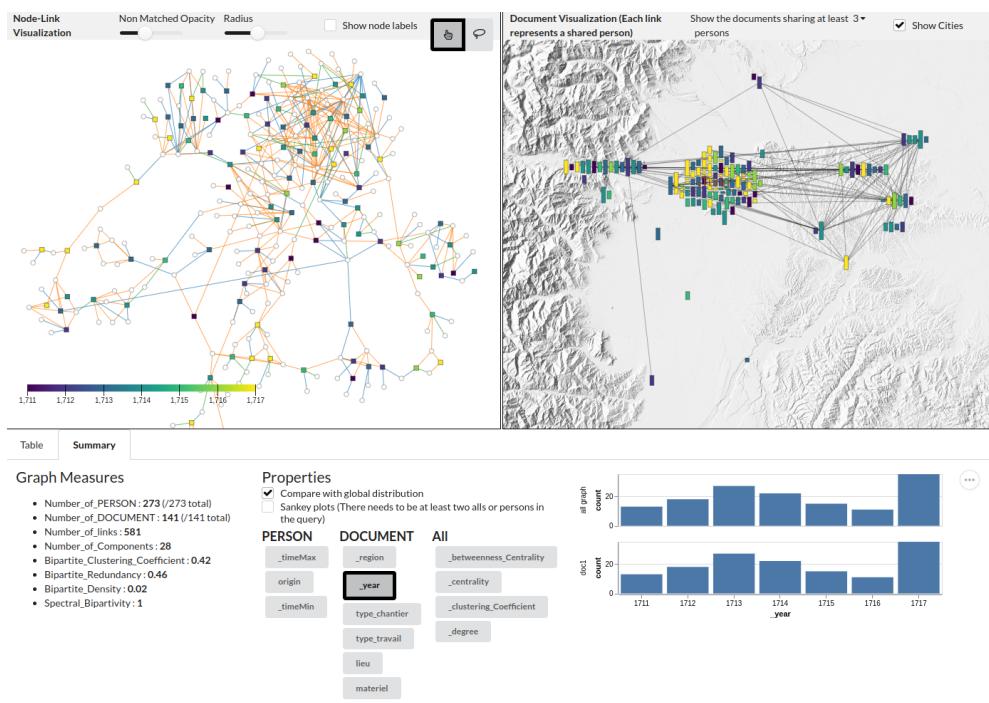


Figure 4.2 – ComBiNet interface with the dataset of collaboration #1. The user selected the year attribute, showing the distribution of document years with a histogram (bottom right), and coloring the documents node on the bipartite view (left) and map view simultaneously (right).

4.4.2 . Query Panel

The query panel allows to rapidly build queries visually, with topological and attribute constraints. The visualization of the query is synchronized with the Cypher query sent to the database. Modifying one representation will update the other, allowing users to build a query visually and refine it in Cypher when appropriate. Experts users who know the Cypher language can also start to construct their query textually and modify it visually later on. In this section, we describe all the features and interactions allowing ComBiNet to build a query and illustrate them

with questions 2 and 6 of the use case #1. Our collaborator wants to *find the persons who are mutually Guarantor to each other in separate contracts* (6) and to know *how Torino and Torino's surroundings differ according to their contracts* ?.

V6 : Node-Link Dynamic Query

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

Node Creation : In *addition* mode, clicking on an empty area creates a new node.

The node will be of the selected type from the legend on the right (Person, Document, or Any).

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

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

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

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

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

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

To find persons who are mutually guarantors in our collaboration #1, we first create one person and two documents using the addition mode and by clicking on the canvas. We then link the person node to the first document with a link that is not typed (Figure 4.3 left), and link it to the second document with a Guarantor

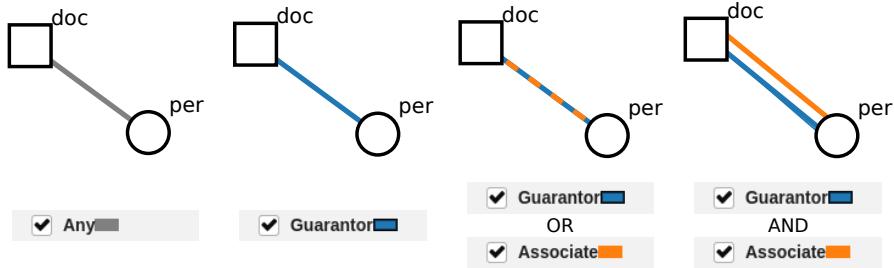


Figure 4.3 – All link creation possibilities : Any link type (left), one selected link type, here guarantor (middle left), the union of several link types (middle right), several links with different types (right)

link (Figure 4.3 middle left). We then create a second person node and link it to the two documents with opposite link types. The resulting visual query is presented in Figure 4.4 (a). To answer the second question, we can simply start to request all the links in the graph, no matter the type, as shown in Figure 4.4 (b). The database will then return all the links in the graph with their attached nodes.

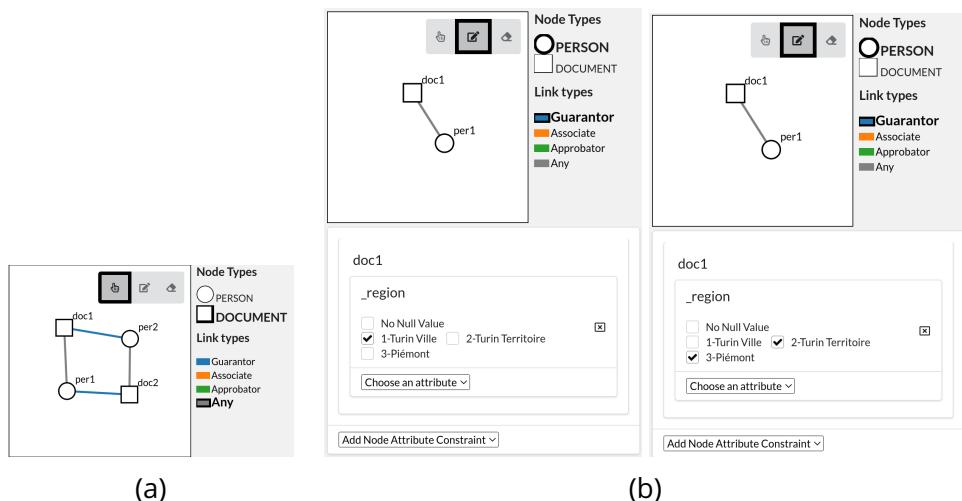


Figure 4.4 – Visual queries created to answer questions 2 and 6 of our collaboration #1. (a) The visual query retrieves individuals who are mutually guarantor to each other in separate construction contracts. (b) The two visual queries retrieve the documents—along with the signatories—of *Turin* (left)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

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

nominal, as in the original dynamic queries [?] :

1. **Numeric constraints** are modeled as range sliders, allowing to select a lower and upper bound to the filter.
2. **Categorical constraints** are modeled as a set of checkboxes. Each possible value has a corresponding checkbox.
3. **Nominal constraints** are modeled as text input, where the user can write any desired value. All the possible values are shown at the same time and filtered as the user writes.

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

The figure displays three filter widgets:

- region**: A categorical filter with checkboxes. It includes options for "No Null Value", "1-Turin Ville" (which is checked), "2-Turin Territoire", and "3-Piémont".
- id**: A nominal filter with a text input field containing "No Null Value".
- date_year**: A numerical filter represented by a horizontal slider with two circular handles, showing a range from 1712 to 1714.

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

To answer our collaborator's second question (*how do Torino and Torino's surroundings differ according to their contracts ?*), we first want to filter the documents which are located in Torino (*Turin* in French). For this, we start by selecting the whole dataset by linking a person and document node with *any* link. Then, we select the id *doc1* of the document of our visual node-link query, and the *region* attribute. It will initialize a categorical widget including all the values found in the dataset for this attribute with associated checkboxes. We check the region of interest “1-Turin Ville” to select all the documents from this region. The first widget of

Figure 4.5 illustrates the created constraint. To select the documents of Torino's surroundings, we can simply uncheck the “1-Turin Ville” value for the *region* attribute and check the two other values “2-Turin Territoire” and “3-Piemont” which are areas corresponding to the surroundings of Torino. Both queries are represented in Figure 4.4 (b).

V8 : Cypher Editor Users can build or modify a query using the Cypher query language, with the Cypher text editor. This allows users to start creating a query visually and refining it by text for complex constraints which can not be represented by a visual form easily. The editor supports autocompletion to e.g., help to discover and spell the attribute names. The visual and textual representations are synchronized, meaning that changing one will update the other and update the results in the visualizations.

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

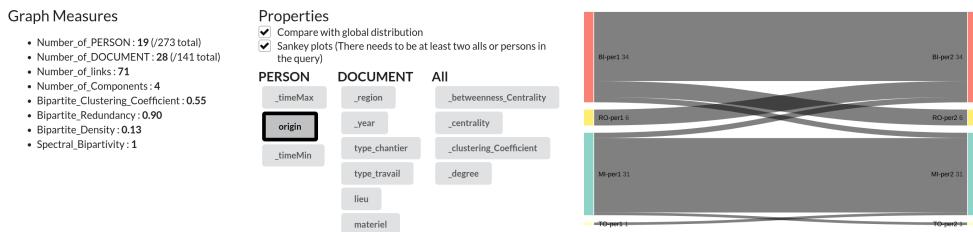
Attributes Visualization When users select an attribute in the attributes view (V5), its distribution is visualized for the queried entities and the whole network. However, these plots show the aggregated values and we lose the potential value transitions between the query nodes. For example, Figure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

Persons Documents Occurrences

CSV Copy Unselect Resize Search:

Option	doc1	doc2	I	I3	I5	I6	per1	per2
DOCUMENT (a25)	DOCUMENT (a4)	Guarantor	Associate	Guarantor	Guarantor	Guarantor	Bello P.(p27)	Gillonio G.P.(p175)
DOCUMENT (a5)	DOCUMENT (a4)	Guarantor	Guarantor	Guarantor	Guarantor	Guarantor	Bello P.(p27)	Gillonio G.P.(p175)
DOCUMENT (a5)	DOCUMENT (a4)	Guarantor	Guarantor	Guarantor	Guarantor	Guarantor	Gillonio G.P.(p175)	Bello P.(p27)
DOCUMENT (a25)	DOCUMENT (a5)	Guarantor	Associate	Guarantor	Guarantor	Guarantor	Bello P.(p27)	Gillonio G.P.(p175)
DOCUMENT (a4)	DOCUMENT (a5)	Guarantor	Guarantor	Guarantor	Guarantor	Guarantor	Bello P.(p27)	Gillonio G.P.(p175)
DOCUMENT (a4)	DOCUMENT (a5)	Guarantor	Guarantor	Guarantor	Guarantor	Guarantor	Gillonio G.P.(p175)	Bello P.(p27)

(a)



(b)

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

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

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

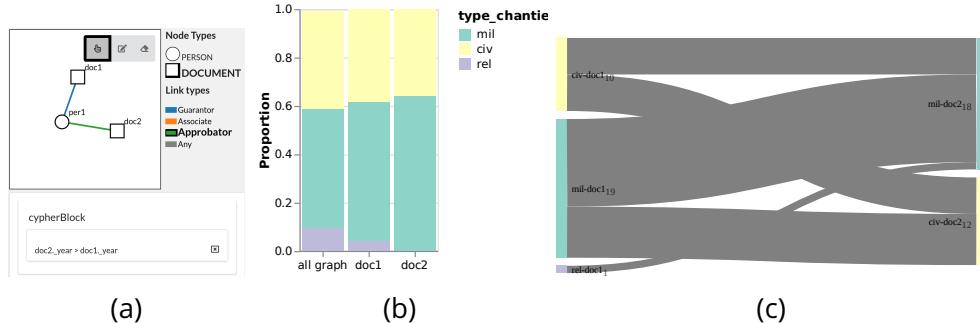


Figure 4.7 – Two ways of showing the distribution of “type chantier” (type of works), a categorical attribute with three possible values “religious”, “military”, and “civilian”. (a) A query matching the contracts made by the same person (*per1*) as an “approbator” (green link to *doc2*) after being a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

to revisit and refine explorations. Figure 4.8 shows the provenance tree made to answer question 2, split into 2 branches, with the tooltip showing one of the node query state.

4.4.3 . Comparison

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

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

For the second example from Table 4.1, we want to compare the works in Torino with the ones in Torino surrounding. Since we previously constructed the query returning all the contracts from *Turin* with the mentioned people, we can return to this point in the provenance tree, and change the constraint of the *region* attribute from *1-Turin Ville* to *2-Turin Territoire* and *3-Piemont* using the checkbox to get the two queries we want to compare. They are shown in Figure 4.4. The user can then rename the provenance tree nodes with explicit names such as “Torino”

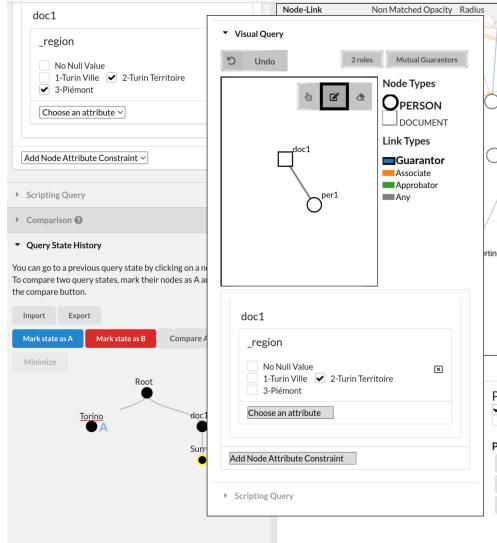


Figure 4.8 – Provenance tree to answer question 2 of collaboration #1 : left branch leads to Torino documents (the node is labeled as A) while right branch leads to surrounding documents (the node is labeled as B). The user hovers over one node, revealing a tooltip that shows the visualization of the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

and “Surroundings”, and mark them as A and B using the appropriate buttons. Clicking on the “Compare State A and B” will make the interface compare the two query results.

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

Check la comparaison

comparaison de torino et territoire sur le graphe

Attribute-Based Comparison The comparison of one or several attribute distributions between (A) and (B) is also useful for answering the historical questions of our users. In the attribute view (V5) of the results panel, hovering or clicking on an attribute name will show the distribution of this attribute in four contexts : the nodes of the whole graph, the queries (A), (B), and the currently selected Boolean operator (e.g., intersection or union) if one is selected. This allows users to compare attribute distributions between several subsets of interest (T3.2). For example, we

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

The

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

can compare the attributes between the contracts of Torino and the ones of its surroundings. We can also compare the persons who worked in Torino, in Torino's close territory, and in both areas, by selecting the intersection operator. Figure 4.10 illustrates the comparison charts for different attributes. We can see that the types of construction sites differ between the two regions : the city of Torino clearly has a lot of military sites compared to the surroundings of Torino, which has almost none. This is the opposite for the number of religious sites, which are almost all localized in the surroundings of Torino. If we now look at the year distribution of the contracts, we can see a difference in the distributions. The years of Torino's construction contracts were steady between 1711 and 1717 with a little spike in 1713, while the constructions were more scarce in the surroundings before 1716. We can see a big spike in construction in 1717. This is interesting to our users, as it shows the dynamic of the construction in the area : the center of the city started to be constructed before other constructions arose in the surroundings.

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

4.4.4 . Implementation

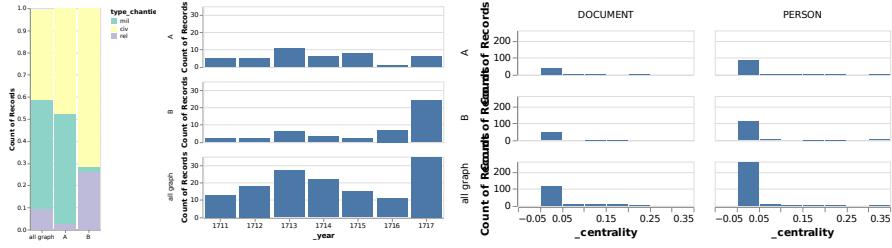


Figure 4.10 – Distribution of the type of constructions, the years, and the centrality for the documents and signatories of Torino (A), Torino surroundings (B), and the whole graph. (top).

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

4.5 . Use Cases

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

4.5.1 . Construction sites in Piedmont (#1)

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

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

We then do the same query for the *Zo* family. We keep the same topological

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

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

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

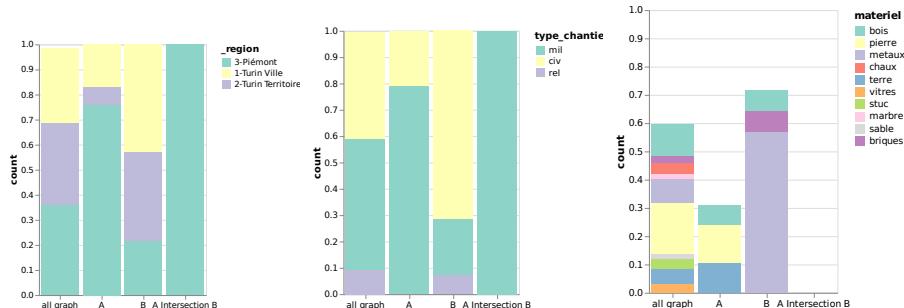


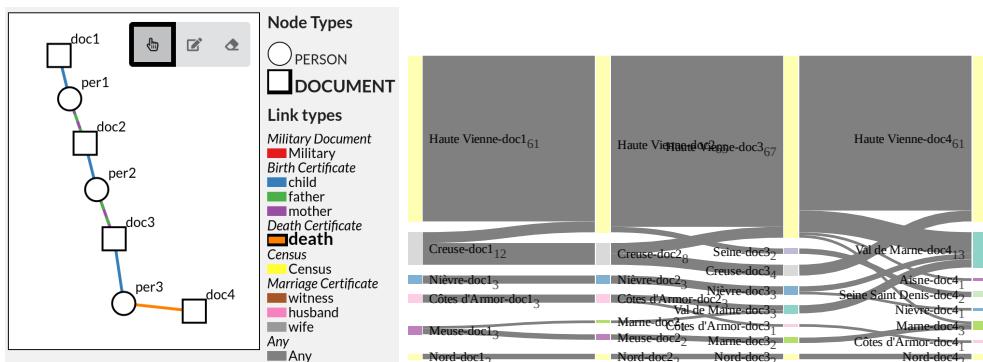
Figure 4.11 – Attributes distributions plots between the whole graph, the *Menafoglio* family (A), the Zo family (B), and $A \cap B$, for the *region*, *type_chantier*, *material* type.

4.5.2 . French Genealogy (#2)

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

To find patterns of migrations within families, we first make a query representing a simple family by linking a person node to a birth event, connected to the parents using a link of *father* or *mother* type. We repeat the process to the new

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



(a) Visual query to find all 3-generation families (b) Sankey diagram showing the birth and death places of people across generations

Figure 4.12 – Migrations across departments over three generations

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

4.5.3 . Sociology thesis in France

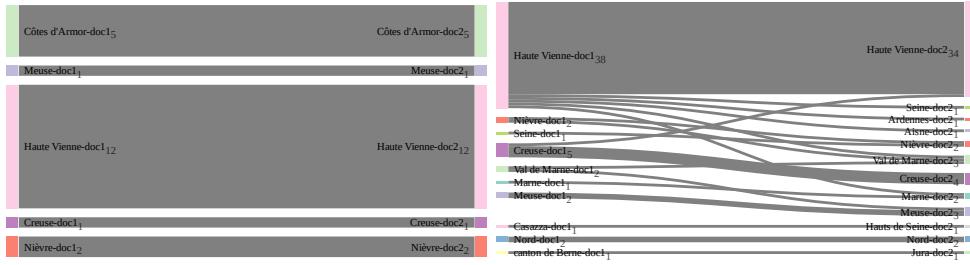


Figure 4.13 – Sankey diagrams showing the migration of people in the 18th and 19th centuries, extracted from their birth and death places.

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

Let's now try to answer an interesting question : "Do referees and jury presidents often ask thesis directors to be referees and jury presidents in their turn of another thesis where they are directors ?". For this, we can construct a visual query representing this pattern by creating two person nodes and two document nodes, and by connecting them with two president links and two referee or jury director links in a symmetrical way, as shown in Figure 4.15 (right). The occurrence table tells us that this pattern has been found 76 times in the network, meaning that this is a recurrent behavior. We are now interested in characterizing the thesis occurring in this pattern, by their regions. We can look at the *city* attribute distribution for

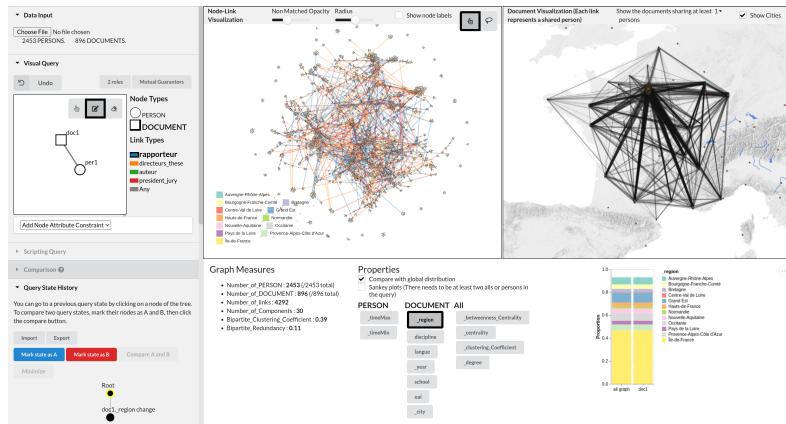


Figure 4.14 – ComBiNet used for exploring theses of sociology defended in France between 2016 and 2021. The bipartite and map views show an overview of two visions of the network. The user selects the *region* attribute, showing the geographical distribution of the defended theses.

this thesis by selecting it in the attribute view as shown in Figure 4.15 (bottom right). We can first see on the map that this pattern occurs mainly in the biggest cities of the country. By selecting the Sankey view option, we can investigate if this pattern occurs between thesis defended in different regions or if it occurs mainly in the same ones. We learn that it depends mainly on the regions : in Bourgogne-Franche-Comté 26 out of 29 theses are connected with the thesis of another region. On contrary, in *Occitanie* it is the case for only 4 out of 17. On average, we can see that this pattern occurs a lot for theses of the same region. In *Ile-de-France*, it is the case for around half of the thesis (28/50). This exploratory analysis shows that ComBiNet can be used to explore and gain insight into such datasets.

4.6 . Formative Usability Study

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

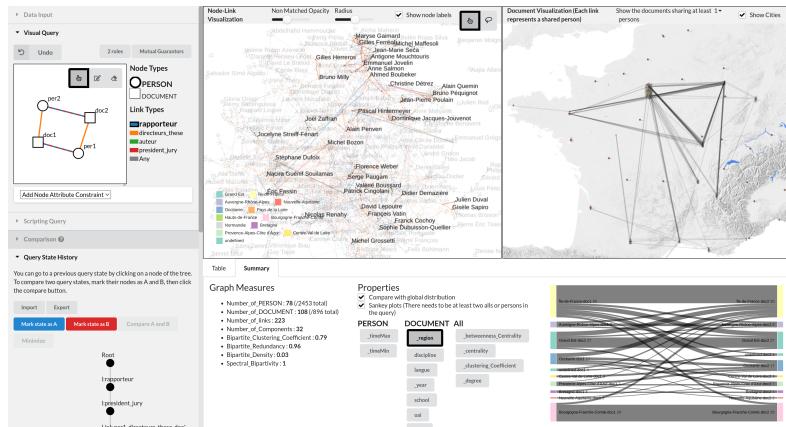


Figure 4.15 – Sociology thesis dataset explored with ComBiNet. The user constructed a visual query to see if there are symmetrical relationships between thesis directors and referees (or jury directors). The *region* attribute is selected with the Sankey option, letting the user see if there are correlations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

nodes across the two views and the tables, and the undo feature for visual queries. At the final meetings, the three users were able to perform exploration, queries, and comparisons to answer socio-historical questions by themselves.

4.6.1 . Feedback

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

4.7 . Discussion

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

lue. The visual query system could be extended by introducing more complex time constraints capabilities, such as in [?].

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

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

4.8 . Conclusion and Future Work

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

By specifying a unifying data model and novel high-level visual and interactive tools for comparing topology, attributes, and time, social scientists were able to

clean their data more easily by finding errors and inconsistencies by exploring the network and querying errors-induced patterns. Thanks to the document-centered model, it was easy for them to trace back the errors and inconsistencies to the sources for corrections. With the same representation, they were able to operate explorations and analyses using complex interactions implemented in ComBiNet such as coordinated views, visual querying, and comparison mechanisms.

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

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

For future work, ComBiNet could be extended to support more SNA measures and computations such as clustering ; it would create a new attribute containing a cluster identifier. The interface currently proposes two layouts based on the topology and the geolocations of the entities. Providing more layout options could be interesting, especially one to highlight better the time, similar to the PAOH-vis technique [70]. Finally, the interface currently lets social scientists build their queries to answer questions they have about their data. In the future, the system could make suggestions on the query construction process with a mixed-initiative perspective, to guide users towards frequent subgraphs in the data which could be interesting to investigate.

5 - PK-Clustering

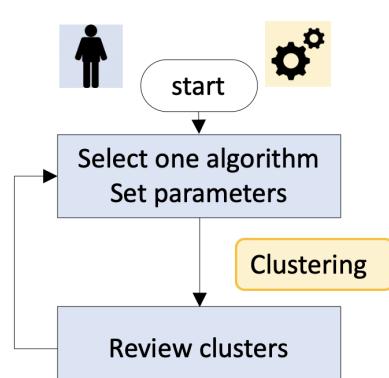
5.1 . Context

The goal of this work is to help social scientists, such as historians and sociologists, create meaningful clusters from social networks they study. In contrast to the belief that most data is easily available on the Web, as of today, most social scientists spend a long time collecting data, to construct social networks, based on documents or surveys, in order to create and carefully validate medium-sized networks (50–500 vertices). Before the start of the cluster analysis a great deal of effort goes into analysing other data and gathering knowledge (which we call prior knowledge in the rest of the paper). Social scientists study in great details the network entities (most of the time people), and the social ties they weave together, as it is the unit brick with which they can make historical or social hypothesis and conclusions. When the network is small, less than 30–50 nodes, it is possible to remember most of the relations and persons and visualization directly helps to show groups, hubs, disconnected entities, outliers, and other interpretable motifs. When the network grows larger, with hundred entities or millions of them, it becomes impossible to perform the visual analysis only at the entity level. The graph has to be summarized, and typically social scientists want to organize it in social *communities*. A large number of algorithms are available today to compute *clusters* of entities from a graph, with the assumption that the computed clusters represent faithfully the social communities. However, most social scientists are not familiar with all of the available algorithms and are challenged to choose which algorithm to run, with which parameters, and how to reconcile the computed clusters with their prior knowledge. Furthermore, the clusters computed by the algorithms do not always align with the concept of community from the social scientists.

Typically, social scientists select an analysis tool based on their familiarity with the tool and the level of local or online support they can access. Therefore, they most often use popular systems such as R [?], Gephi [4], Python with NetworkX [?], or Pajek [?]. To compute clusters, they follow a strained process : they select and run algorithms provided in the tool and then try to make sense of the results (see Figure 5.1). When they are not satisfied or unsure, they iteratively tweak the parameters of the algorithms at hand, run them again and hope to get results more aligned with their prior knowledge. This analysis process is unsatisfactory for three main reasons :

1. it forces them to try a sometimes large number of black-box algorithms one by one, tweaking parameters that often do not make

Traditional Clustering



- sense to them ;
2. even when a parameter makes sense to them, such as the number of clusters to compute, k in k -means clustering, they have no clue of what value would generate good results, and are left with trial and error ;
 3. even if they could painstakingly evaluate the results of all clustering algorithms according to their prior knowledge, no existing system allows users to do so easily, leading users to give up and blindly accept the results of one of the first algorithms they try.

Those complaints have been heard repetitively during the decades our team has worked with social scientists.

Moreover, clustering is an ill-defined problem : for one dataset, there is no ground truth, and several partitions can be considered good according to the metric chosen to evaluate the result [?]. In a Social Sciences setting, this means, for example, that the same social network could be clustered to find families, friend groups, or business relationships. One partition is not better than the other : it depends on the purpose of the analysis. This problem increases the need for interactive tools, which let the user specify which type of partition is expected.

To address those issues we propose a novel approach, called PK-clustering, which allows social scientists to iteratively construct and validate clusters using both their *prior knowledge* and consensus among clustering algorithms. A prototype system illustrates such approach.

The proposed approach includes three main steps (see Figure 5.2) :

1. *Specify Prior Knowledge (PK)*. Users introduce their prior knowledge of the domain by defining partial clusters. The tool then runs all available clustering algorithms.
2. *Consolidate expanded PK clusters*. Users review the list of algorithms, ranked according to how well they match the prior knowledge. They compare results and consensus, then accept or ignore suggestions to expand the prior knowledge clusters
3. *Consolidate extra clusters*. The tool suggests extra clusters on unassigned nodes. The user reviews consensus on each proposed cluster, then accepts or rejects suggestions.

The output of the process is, using a direct quote from a social scientist providing feedback on the prototype : "a clustering that is supported by algorithms and

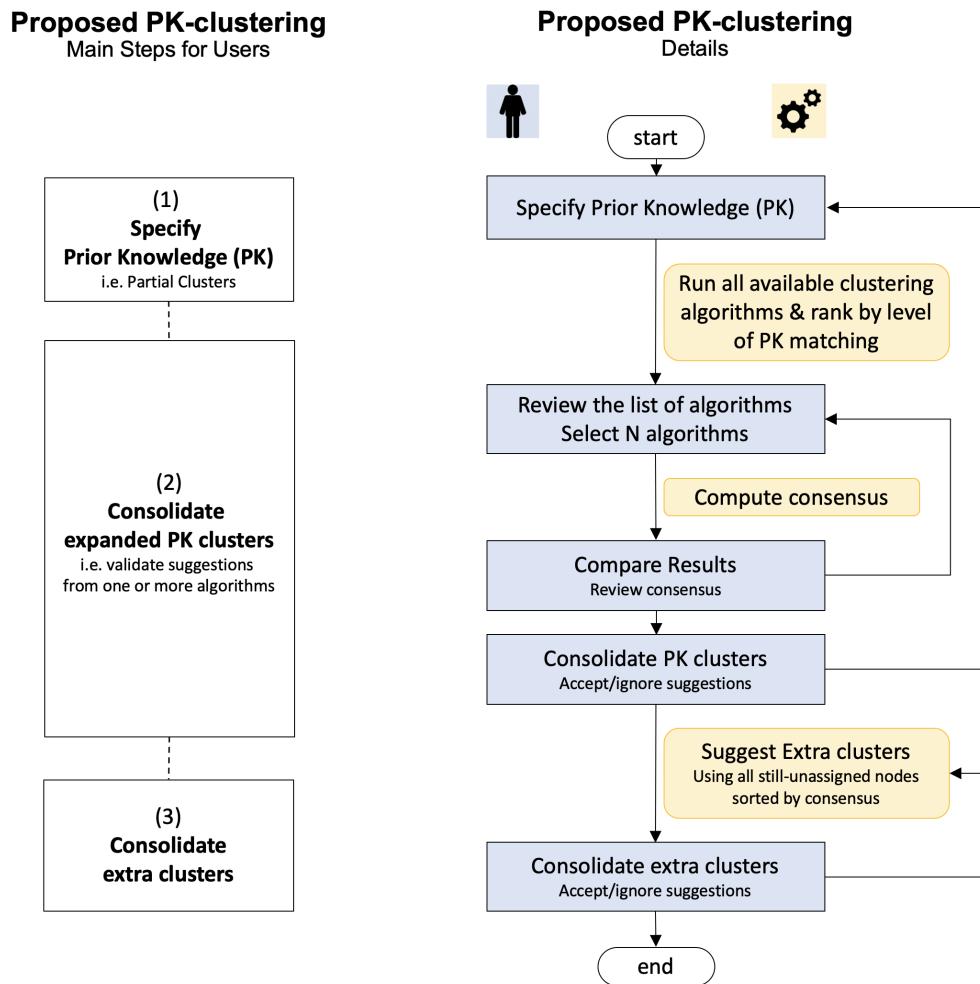


Figure 5.2 – PK-clustering. The output is a clustering supported by algorithms and validated (fully or partially) according to the user's Prior Knowledge.

validated, fully or partially, by social scientists according to their prior knowledge".

According to the need to combine data mining with visualizations [?] and inspired by the idea of letting the user collaborate with the machine to reach specific goals [?], the proposed approach follows a user-initiated mixed-initiative [?] visual analytics process.

In our case, users focus on the results that expand on their prior knowledge, filter-out the most implausible results, but can readjust when they realize that several algorithms are consensual despite not matching the prior knowledge (hinting at other possible meaningful structures). Our mixed-initiative approach allows social scientists to seed the clustering process with a small set of well-known entities that will be quickly and robustly expanded into meaningful clusters (details in subsection 5.3.1).

Contrary to a current trend [?], we do not aim to improve the interpretability of algorithms but to improve the interpretation of the results of black-box algorithms in light of prior knowledge, provided by the user. Every day, we use complex mechanisms that we do not fully understand, like motorbikes, cars or electric vehicles using various kinds of engines, shifts, and gears, but we are still able to choose which one best fit our needs according to their external utility and not by understanding their complex internal machinery. In addition, it is usually more important to social scientists to find an algorithm that provides useful results than to understand why another algorithm failed to do so.

The main contributions of this article are :

1. a new interactive clustering approach ;
2. a prototype (shown in ??) implementing PK-clustering with 11 clustering algorithms of different families applied with different parameters configurations ;
3. two case studies.

5.2 . Related Work

Our approach relies on several families of clustering methods and the visualization and exploration of their results. We first describe a brief overview of clustering for graphs, as well as semi-supervised methods, then several works in the literature related to visual analytics : interactive clustering, groups in networks and ensemble cluster visualization.

5.2.1 . Graph Clustering

One of the main properties of social networks is their community structure [?] that reveals group relationships between nodes, known as communities or clusters, having higher density of edges than the rest of the graph. Similar characteristics or roles are often shared between nodes of the same community. In social networks, a community can mean a lot of things like families, workgroups, or friend groups. There is abundant and growing literature on clustering methods to find these communities for social networks. The majority of the research is made only

on topological algorithms, i.e., algorithms which use only the structure of the network to find clusters. [?] proposes a description and a classification of various algorithms, such as divisive, spectral and dynamic algorithms, or methods, such as modularity-based, statistical inference, to cite a few. In contrast, many multi-dimensional clustering algorithms use a distance function as parameter, but graph clustering algorithms mainly rely on the structure of the graph instead.

Even if the majority of studies are based on simple graphs, real-word phenomena are often best modeled with bipartite graphs, also known as 2-mode networks. It is the case for social scientists, who often build their networks from raw documents containing mentions of people. In that case, it is more straightforward to model the persons as one set of nodes, the documents as the other one, and linking an individual to a document if the individual is mentioned in it. This is one of the reasons some research is made on bipartite graph community detection [?].

Moreover, recent new approaches try to use the attributes of the nodes [?] and the dynamic aspect of the networks [?] to find more relevant communities. Some toolkits offer a large number of algorithms ; for example, the Community Discovery Library (CDLIB) [?] implements more than 30 clustering methods with variations inspired by 67 references.

5.2.2 . Semi-supervised Clustering

In semi-supervised clustering the user integrates the data mining task with additional information to improve the clustering quality in terms of minimizing the error in assigning the cluster to each data of interest.

Semi-supervised clustering can be divided into constraint-based and seed-based clustering. The former includes must-link (ML) and cannot-link (CL) constraints [?, ?]. $ML(x, y)$ indicates that given two items x and y , they must belong to the same cluster, while $CL(x, y)$ means that x and y must belong to different clusters.

Seed-based clustering requires a small set of seeds to improve the clustering quality. Several works addressing seed-based clustering have been proposed in the literature, such as : k -means [?], Fuzzy-CMeans [?], hierarchical clustering [?], Density-Based Clustering [?], and graph-based clustering [?]. Shang et al. [?] use a seeding then expanding scheme to discover communities in a network. Their clustering method considers edges as documents and nodes as terms.

Swant and Prabukumar [?] review graph-based semi-supervised learning methods in the domain of hyperspectral images. Nodes of the graph represent items that may be labeled, while the edges are used to specify the similarity among the items. The technique classifies unlabelled items according to the weighted distance from the labeled items.

5.2.3 . Mixed-Initiative Systems and Interactive Clustering

Introduced by Horviz [?], mixed-initiative systems are “interfaces that enable users and intelligent agents to collaborate efficiently”. Several Visual Analytics systems are based on mixed-initiative interactions, e.g. [?, ?, ?, ?], in particular the

interactive clustering systems.

PK-Clustering is an interactive clustering system. A review by Bae et al [?] shares our concerns : “Real-world data may contain different plausible groupings, and a fully unsupervised clustering has no way to establish a grouping that suits the user’s needs, because this requires external domain knowledge.” Interactive clustering systems aim at producing visual tools that let users interact and compare several clustering results with their parameter spaces, making it easier to find a satisfactory algorithm for a particular application. Several such systems exist (e.g. [?, ?]) but few deal with graph data. These systems adapt one algorithm to become interactive using some type of constraints. Instead, our approach applies ML/CL constraints on a wide variety of existing algorithms, providing richer algorithms and control than the reviewed systems.

5.2.4 . Groups in Network Visualization

To assess the quality of clusters in graphs, the clusters should be visualized. A state of the art report (STAR) on the visualization of group structures in graphs is proposed by Vehlow et al. [?]. Several strategies exist to display group information on top of node-link diagrams. Jianu et al. evaluated four of them : node coloring, LineSets, GMap and BubbleSets [?]. They show that BubbleSets is the best technique for tasks requiring group membership assessment. But, displaying group information on a node-link diagram can reduce the accuracy by up to 25 percent when solving network tasks. Another finding is that the use of GMap of prominent group labels improves memorability. Saket et al. evaluated the same four strategies [?], using new tasks assessing group-level understanding.

Holten [?] proposes edge bundling on compound graphs. He bundles together adjacent edges, making explicit group relationships at the cost of losing the detailed relationships. A good example of manual grouping and tagging is SandBox, which allows users to organize bits of information and their provenance in order to conduct an analysis of competing hypotheses [?]. A lot of work has also been done on the visualization of categorical variable in tabular data [?, ?], which is similar to the notion of groups in networks.

5.2.5 . Ensemble Clustering

In the context of machine learning, an ensemble can be defined as “a system that is constructed with a set of individual models working in parallel whose outputs are combined with a decision fusion strategy to produce a single answer for a given problem” [?]. Several strategies exist for combining multiple partitions of items in a clustering setting [?]. Concerning visualization research, Kumpf et al. [?] consider ensemble visualization as a sub-field of uncertainty visualization, for which some surveys exist [?, ?]. They describe a novel interactive visual interface that shows the structural fluctuation of identified clusters, together with the discrepancy in cluster membership for specific instances and the incertitude in discovered trends of spatial locations. They aim at identifying ensemble members that can be considered similar

and propose three different compact representation of clustering memberships for each member. Our system provides a consensus based interactive strategy that takes into account user's prior knowledge instead of relying on mathematically defined optimal assignments only.

5.2.6 . Summary

The community detection problem in graphs has been studied in a lot of different settings. We can classify it this way from the user perspective :

Standard clustering. One algorithm is picked with a set of parameters and the user check if the results are consistent with his prior knowledge, which is not represented in the process.

Ensemble clustering. Many algorithms run with potentially many parameters, and a final partition is obtained by trying to merge optimally the partitions. At the end of the process, one clustering is given to the user who has to check if it is consistent with the prior knowledge, which is not used either.

Semi-supervised clustering. The user provides the prior knowledge and lets the algorithm propose a final solution using this information in its computation. The results should be good by design, regarding the knowledge of the user.

The aim of our proposed framework is to combine these three approaches, to integrate the user in the analysis loop and allow him to have a better impact on the final community detection result.

5.3 . PK-clustering

We present a new approach, inspired by the three types of clustering methods described in subsection 5.2.6 : Standard clustering, Ensemble clustering and Semi-supervised clustering. It runs a set of algorithms, then highlights those that best match the prior knowledge provided by the domain expert. The user then reviews and compares the results of the selected algorithms, in order to consolidate a satisfactory and consensual partition.

PK-clustering is not tied to any specific graph representation technique and could be used to augment any of them. Our prototype is implemented in the PAOHVis tool [?] which illustrates how users can view their networks as PAOH (Parallel Aggregated Ordered Hypergraph) or traditional Node Link diagrams. PK-clustering relies heavily on having a list of nodes, so the PAOH representation is naturally adapted to PK-clustering, and will be used in all the figures.

After a general overview of the process, we describe each step in more details, illustrated with screen samples taken during the analysis of a small fictitious network.

5.3.1 . Overview

In PK-clustering the user and the system take turn to construct and validate clusters. The process involves three main steps, each with several activities (see

Figure 5.2. The blue boxes describe the user activities while the yellow boxes describe the system activities.) After loading the dataset, the process is as follows :

(1) Specify Prior Knowledge (PK).

1. The domain experts interactively specify the PK by defining groups, i.e., naming groups and assigning entities to them. Typically, an expert would assign a few items (1-3) to a few groups (2-5), thus creating a set of partial clusters.
2. All available clustering algorithms are run. Algorithm parameters (e.g., number of clusters) may also be varied manually or automatically using a grid search or a more sophisticated strategy, resulting in additional results. Depending on the type of algorithm, topology and/or data attributes are used. The specified PK is used by the semi-supervised algorithms, which are the only ones able to use it.

(2) Consolidate expanded PK clusters.

3. Users review the ranked list of algorithms. They can see if the algorithm results match the PK completely, partially or not at all. Information about the number of clusters generated by each algorithm is also provided. Users select the set of N algorithms they think are the most appropriate.
4. The consensus between the selected algorithms is computed and visualized next to the graph visualization (in the PAOHVis display in our prototype)
5. Users review and compare the suggestions made by the algorithms to expand the PK-groups into larger clusters and examine consensus between algorithms.
6. Users accept, ignore, or change the cluster assignments. This consolidation phase is crucial, as users take into account their knowledge of the data, the graph visualization, and the results of the clustering algorithms to make their choices.

(3) Consolidate extra clusters.

7. The system proposes extra clusters using nodes that have not been consolidated yet and remain unassigned. Users can select any algorithm and see the extra clusters it suggests.
8. For each proposed cluster, users can see if other algorithms have found similar clusters, and then consolidate again by accepting, ignoring, or changing the suggestions for all the nodes in the proposed cluster. This step is repeated with other clusters until the user is satisfied.

/devAt any point users can go back, select different algorithms, or even change the PK specification to add new partial clusters. Users can also opt not to specify any PK partial clusters at all, and accept all consensual suggestions without reviewing them in details. This gives users control over how much they want to be involved in the process. Similarly, users are not required to assign every single node to a cluster.

By specifying the PK in the first phase, before running the algorithms, users avoid being influenced by the first clustering results they en-

Prior Knowledge specification

Prior Knowledge		Create Group	Delete
<input checked="" type="checkbox"/>	A		
	Elise		
	Jacques		
<input checked="" type="checkbox"/>	B		
	Hubert		
	Vallet		

Figure 5.3 – Prior Knowledge specification, the user defi-

counter. The process leads to algorithms whose results match the PK, but it also allows to review results that contradict it.

We believe that PK-clustering addresses the important problems identified in the introduction : it helps users decide which algorithm(s) to use, facilitates the review of the results taking into consideration both the consensus between algorithms and the knowledge users have of their data. We will now review each step in more details.

5.3.2 . Specification of Prior Knowledge

We ask users to represent prior knowledge as a set of groups. Each group contains the node(s) that the expert is confident belong to the defined group. In the case of Figure 5.3, each of the two prior knowledge groups contains two nodes, and it specifies that the user is expecting to see at least two clusters, with the first two people in a blue cluster A, and the other two in a red cluster B. This representation expresses *must-link* and *cannot-link* constraints described in subsection 5.2.2 in a simple visual and compact form. It is not required to specify all binary constraints because the information is derived from the prior knowledge groups.

5.3.3 . Running the Clustering Algorithms

Our prototype includes 11 algorithms taken from three families :

Attribute based algorithms. Graph nodes can have intrinsic or computed attributes that can be used for grouping, such as gender, family name and age. Some community detection algorithms use those attributes alone or together with the topology to partition the graph. A clustering algorithm considers attributes according to their type. For categorical attributes (e.g., male / female) it finds matching attributes and merges them if necessary. For numerical attributes (e.g., income) the algorithm seeks to define intervals which can be adjusted for propagating clusters. Algorithms in this family can also use multiple attributes together.

Topology based algorithms. Most of the clustering algorithms consider only the graph topology [?] and try to optimize a topological measure such as *modularity* [?]. Those algorithms only use the connections between the people to find groups. Their aim is to find groups of nodes such that the density of edges is higher between the nodes of one group than between the group and the rest of the graph.

Propagation / Learning based algorithms. Semi-supervised machine learning algorithms learn from an incomplete labeling of data and use it to classify the rest of the data. They represent a class of machine learning methods, also called label propagation methods, which can take into account users' Prior Knowledge groups in its clusters computation. By design, this type of algorithms will always provide

a perfect match with the Prior Knowledge, even if the Prior Knowledge makes no sense.

Our prototype implements 2 attribute based algorithms (one for numerical attributes and the other for categorical attributes), 7 topology based algorithms and 2 propagation based. Since we often deal with hypergraphs 2 of the topology-based algorithms are bipartite node clustering algorithms : Spectral-co-Clustering [?] and Bipartite Modularity Optimisation. Since the majority of community detection algorithms are for unipartite graphs, we perform a projection into a one-mode network [?]. Basically, each pair of nodes which are in the same hyperedge are connected together in the resulting graph, with a weight being the number of shared hyperedges [?].

Some algorithms require parameters to be specified. We do not force the user to specify values for all the parameters, when possible, we infer them from the PK-groups. For instance, instead of using an arbitrary default for the number of expected clusters k in k -means clustering, we run the algorithm several times with a value of k from the number of specified PK-groups to this number plus two. Therefore, our implementation computes a total of 15 clustering algorithms (11+4). The strategy of using several parameter combinations for the same algorithm is often used in ensemble clustering to increase the number of different clusterings. However, the number of parameter combinations can be extremely high. The research field of *visual parameter space exploration* (see e.g., [?]) is devoted to exploring this space of parameter values in a sensible way ; we currently address the problem only for simple cases.

Once all the algorithms finish the computation, we try to match the resulting partitions with the PK and rank the algorithms by how interesting their results are likely to be for the user.

5.3.4 . Matching Clustering Results and Prior Knowledge

Once a clustering is computed, we want to know how well it is compatible to the PK, and if possible, match every PK-group with a specific cluster. We use the *edit distance* to measure this matching, as its computation allows us to directly link each PK-group to a specific cluster. Given two partitions, the edit distance is the number of single transitions to transform the first partition into the second one. For example, the edit distance between the two partitions of 4 nodes $P_1 = \{\{1, 2, 3\}, \{4\}\}$ and $P_2 = \{\{1, 2\}, \{3, 4\}\}$ is 1 because moving the node 3 from the first to the second set of P_1 would transform it into P_2 . A clustering can be seen as a partition since every node has a label, but the PK can only be seen as a partial partition because only some nodes are labeled. We say that the edit distance between the PK and a clustering is 0 if every group of the PK is a subset of an exclusive cluster, i.e., if every person of a PK-group is retrieved in the same cluster, with no overlaps. Thus, we define the edit distance as the number of node transitions between the groups of the PK to get to the state where each group is a subset of an exclusive cluster.

To compute the edit distance and the matching, we build a bipartite graph : each meta-node corresponds either to a PK-group, or a cluster. We then link them if they share a node, with a weight equals to the number of shared nodes. Computing the edit distance and producing a matching between the PK-groups and the clusters is then equivalent to the assignment problem, where the goal is to find a maximum-weight matching in the graph. [?].

Once this matching is computed, the total sum of the weights minus the sum of the weight of the matching is equivalent to the number of transitions needed to transform the first partition into the second one (or the PK into a sub-partition where each set is an exclusive subset of the sets of the second partition), i.e., the edit distance.

For example, given a clustering of 12 nodes $N = 1, 2, \dots, 12$, the clusters $C_1 = [1, 2, 3, 4]$, $C_2 = [7, 9, 10, 12]$ and $C_3 = [5, 6, 8, 11]$ and a PK composed of 3 groups $PK_1 = [1, 2]$, $PK_2 = [5]$ and $PK_3 = [3, 7]$, the maximum-weight matching is given by the edges (PK_1, C_1) , (PK_2, C_3) and (PK_3, C_2) . This is illustrated in Figure 5.4. The edges of the matching correspond to the matching between

the PK-groups and the clusters. The edit distance is then equal to the sum of all the weights of the bipartite graph minus the sum of the weights of the maximum matching (in red), thus equaling $5 - 4 = 1$. In other words, we only have to move the node 3 from PK_3 to PK_1 , for every PK-group to be a subset of an unique cluster, with no overlap.

In the end, we hope to find matches linking every PK-group to one specific cluster, with no overlaps. This is not always the case and sometimes two or more PK-groups are subsets of the same cluster. In that case, it is not possible to link all these PK-groups to the same cluster since we want one unique cluster for each group. Thus, we say that the algorithm failed to match the prior knowledge and we do not summarize it visually.

5.3.5 . Ranking the Algorithms

The algorithms are ranked by their degree of matching with the prior knowledge, using the edit distance. We also introduce a *parsimony* criterion if there is a tie between two or more algorithms. The algorithm with the smaller number of other clusters will be shown first, as the results are easier to interpret. Moreover, the number of specified prior knowledge groups is expected to be close to the final number of clusters the user wants to retrieve, as social scientists often have a good knowledge of their data.

To complement the parsimony rule, we also consider that the family of propa-

gation/learning based clustering algorithms is more complex than the two previous families (attribute or topological based clustering), in the sense that they are more difficult to explain. If a simple and a complex algorithm match the prior knowledge, the simpler one is presented first. For example, if grouping by the attribute “profession” provides a perfect match, then it is ranked higher than a propagation based method achieving the same perfect match.

Semi-supervised methods will always provide a perfect match by definition. But if all the other algorithms (topological and attribute based) do not give a match, it means that the PK does not align well with the data. This would signal the user to reconsider his PK or provides more information in the graph.

5.3.6 . Reviewing the Ranked List of Algorithms

Once the clustering algorithms have been matched with the PK, users can review the list of algorithms, ranked by how well their results match the PK. Figure 5.5 shows two modalities to visualize the ranked list (individual nodes, and aggregate representation). We will describe in details the first modality, which shows individual nodes as small colored circles (also used on the left of ??) :

Each row is an algorithm, and the algorithms are grouped by family. On the right of the name of the algorithm we can see a representation of the clusters that best match each of the PK-groups. In Figure 5.5 we first see the cluster which best matches the blue PK-group, and then the cluster which best matches the red PK-group. In each cluster we see colored dots for each person that matches, and dark gray dots with a X for no match. Additional nodes in the cluster are represented as white dots with a number next to it. On the right most we see how many other clusters (if any) have been found by the algorithm - also represented as white dots with a number next to it.

So for example, the second algorithm *fluid_k3* has a blue cluster that matches the blue PK-group plus 1 extra node, a red cluster that matches the red PK-group plus 5 nodes, and one extra cluster. We see that the top four algorithms match the PK perfectly, while the next one *fluid_k4* have a partial match. At the bottom, an algorithm has no match.

The alternate modality of representing the matches (shown at the bottom of Figure 5.5) uses bars to aggregate the nodes and show the proportion of mat-

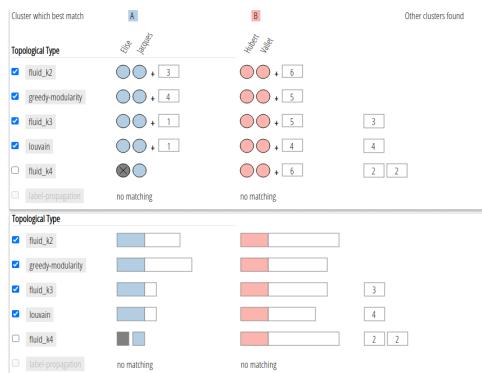


Figure 5.5 – Two different modalities for the ranked list of algorithms. Top : persons are shown as circles. Bottom : aggregated view. Colors indicate the matching group. Gray indicates no match. White indicates extra nodes or clusters.

ching, non-matching and other nodes in each cluster . This is more useful when dealing with bigger graphs, because it allows the user to see the results in a more compact way.

Once users have reviewed the list of algorithms they can review results of a single algorithm, or review and compare the results of all the selected algorithms. By default only the top algorithms are selected for inspection, but users can select any set of algorithms according to different criterion : the *degree of matching* (i.e., they can choose to look at algorithms with no match to challenge their prior knowledge) ; the *algorithm type* (the user may prefer an attribute-based algorithm, rather than one based on topology) ; the *size* of the matched clusters ; or the number and size of *other clusters* found by the algorithm.

PK-Clustering expresses its prior knowledge through *must-link* and *cannot-link* constraints. However, at this stage, the user can decide to use this expressive power as strong constraints—only selecting algorithms that match all of them—or as weak constraints—to explore clustering results that support most or some of them. Our historian colleagues have used both, either to cluster a well-understood dataset with strong constraints or to generate hypotheses on less known ones.

5.3.7 . Reviewing and Consolidating Final Results

To consolidate the final results several approaches are possible. Applying mixed-initiative principles users can rapidly accept labels from a specific algorithm (which is particularly useful for large datasets), or review consensus between selected algorithms then accept only consensual suggestions, or dig in manually to review labels one by one, override labels when appropriate, or leave certain nodes unlabeled. The tool generally guides users to first focus on the PK clusters, then other clusters. The notion of prior knowledge can evolve during the exploration and the process can be iterated from the beginning when new knowledge is gained, thus giving new algorithm matches. Therefore, the approach is not linear but can be iterative.

Reviewing Results of a Single Algorithm

By clicking on an algorithm name the results of that algorithm are displayed in the PAOHVis view (see Figure 5.6). In this view, each line corresponds to a person in the graph, and each vertical line represents an hyperedge connecting them [?], in a way visually similar to the UpSet representation [?] but semantically different. Alternative graph representations are available as well—such as node link diagrams—but the PAOHVis view is well adapted to PK-Clustering.

Names are grouped by the proposed clusters. Clusters that match the prior knowledge are at the top, colored by their respective colors. Black borders around labels highlight nodes that belong to the PK, making them easy to find. All the other (non PK) clusters are initially regrouped in a single group labeled *Others*. A click on the *Others* label expands the group into the additional clusters defined by the selected algorithm. Users can rename the clusters, and change which algorithm

is used for grouping and coloring the nodes.

Comparing Multiple Algorithm Results

From the ranked list of algorithms users can select a set of algorithms and click the large green button to review and compare the selected algorithms in the PAOHVis view (see Figure 5.6 and also ?? for overall context). By default, the PAOHVis view groups the names using the clusters of the 1st algorithm, but on the left of the node names now appears complementary information about the results of all the selected algorithms.

On the far left, the consensus distribution appears as a horizontal stacked bar chart. The size of bar segments corresponds to the number of algorithms that associate the specific node to the cluster having the same color. On the right of the stacked bar chart, first appears the prior knowledge (with square icons). Icons and names of PK nodes have a black border. Further right are shown the individual algorithms' results, represented by diamonds, one for each node and algorithm. When the node is classified in one of the clusters matching a PK-group the diamond is colored with the color of that group.

For each node, the horizontal pattern of colored diamonds quickly tell users if there is agreement among the algorithms. If all algorithms agree the line of diamonds is of a single color. Conversely, if they disagree diamonds will vary in color. If a node does not match any PK-group then no icon is displayed in this phase.

In Figure 5.6 PK_louvain is selected as the base algorithm for the grouping of names in the list. We see that there is very good consensus on the red cluster, but in the blue cluster only 4 out of 7 algorithms see Joseph as belonging to it. Others see him as belonging to the red cluster. In *Others*, 4 algorithms consistently disagree by assigning 3 more nodes to the blue cluster. There are clearly many ways to cluster data, and users must decide the more meaningful one, based on their deep knowledge of the people in the network before validating clusters, possibly by re-reading source documents or gathering more.

Consolidating the prior knowledge clusters

Next, using their knowledge and the consensus of the algorithms, users validate clusters that expand the prior knowledge groups. We call the validated data *consolidated knowledge*. It is kept in an additional column on the right of the algorithms, left of the names. The tool provides several ways to consolidate knowledge and keeps track of the decisions :

Partial Copy. By clicking on one of the icons or dragging the cursor down on a set of icons, users validate the suggestion(s) of an algorithm, adding colored squares in the consolidation column. Once this validation is done, the squares do not change color anymore and represent the user's final decision (unless changed

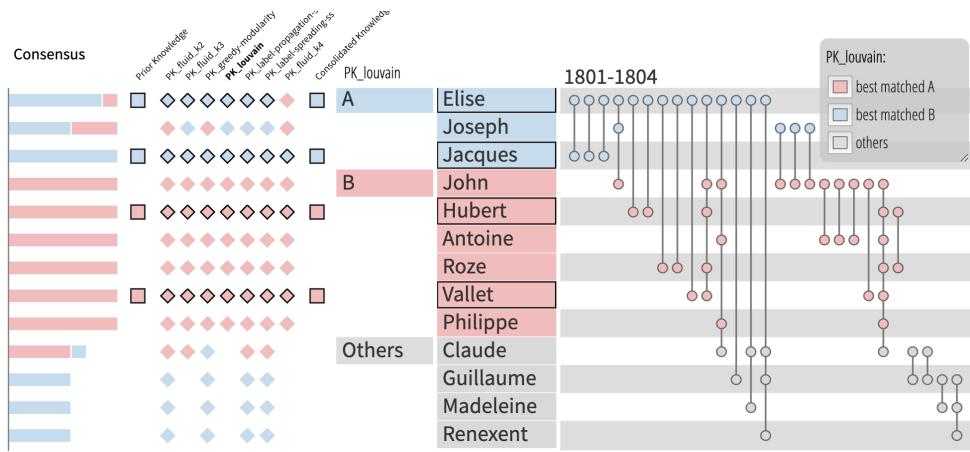


Figure 5.6 – Reviewing and comparing results of multiple algorithms. One algorithm is selected to order the names and group them, but icons show how other algorithms cluster the nodes differently, summarized in the consensus bar on the left.

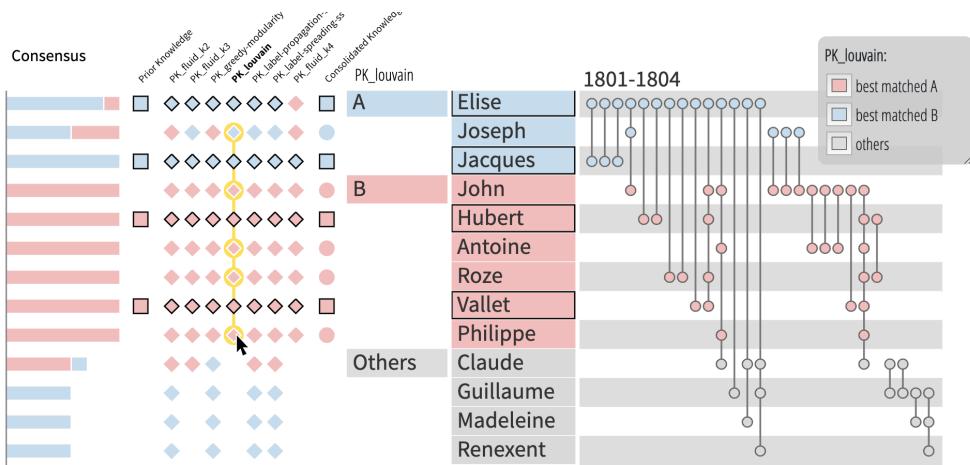


Figure 5.7 – The user quickly drags on consecutive icons (in yellow) representing the suggestions made by one algorithm to validate node clustering. Once the cursor is released the validated nodes appear as squares icons in the Consolidated Knowledge column.

manually again). Figure 5.7 shows how a user drag-selects a set of diamonds in the column PK_fluid_k4. They are connected by a yellow line, which appears while dragging over the icons. When done the status of the nodes in the Consolidated Knowledge column (rightmost) will change to square.

Consensus slider. Users can set the consensus slider to a certain value (for example 4) to automatically select all nodes that have been classified in the same cluster by at least 4 algorithms. While the slider is being manipulated circles appear in the consolidated column. Then users can validate the suggestions by clicking or dragging on the circles, or by using the *consolidate suggestions* button which will validate all suggestions at once. This button is shown in Fig. ???. In summary, diamonds represent suggestions from one algorithm, circles temporary choices, and squares represents the knowledge validated by the user.

Direct tagging. At any time, users can manually overwrite the association of a node to a cluster by right clicking on the node in the consolidated knowledge column and selecting an cluster from a menu. When no clear decision can be made users can leave nodes unassigned, and no shape is displayed in the consolidated knowledge column.

Consolidating extra clusters

The last step of PK-clustering aims to find new clusters for the nodes that have not been validated yet, based on the consensus of the selected algorithms. The suggestions are made from the point of view of one clustering algorithm that the user can change along the process. First, the user selects one algorithm in the PAOHVis view and the nodes are grouped by the clusters found by the algorithm. The PK-clusters are displayed at the top, followed by *Others*, which contains everyone else. When users click on *Others*, the other clusters are displayed ordered by consensus. Since the number of clusters can be high, all new clusters appear in gray to avoid the rainbow effect. A secondary matching process matches the clusters of the current algorithm with those of all the other algorithms, one by one (similar to the matching process described in subsection 5.3.4) . Once the matching is done, the consensus of one cluster is computed as the sum of the cardinalities of the intersections between the cluster and all the other clusters of the other algorithms matched with it, divided by the number of nodes of the cluster.

When users hover over one cluster name, a new color is given to that cluster (e.g., green) and new (green) diamonds appear for each algorithm that match the cluster and for each node that is assigned to the cluster (Figure 5.8). Users can therefore see if the selected cluster is consensual, and with which algorithms. The top part of Figure 5.8 shows the mouse pointer before hovering on the cluster 2. The bottom part shows that hovering the mouse pointer over the cluster 2, it changes to green and several green diamonds appear along three columns.

The evaluation of the best cluster for a node can be done using multiple encodings. The suggested clusters appear into the consensus bar chart, in the set

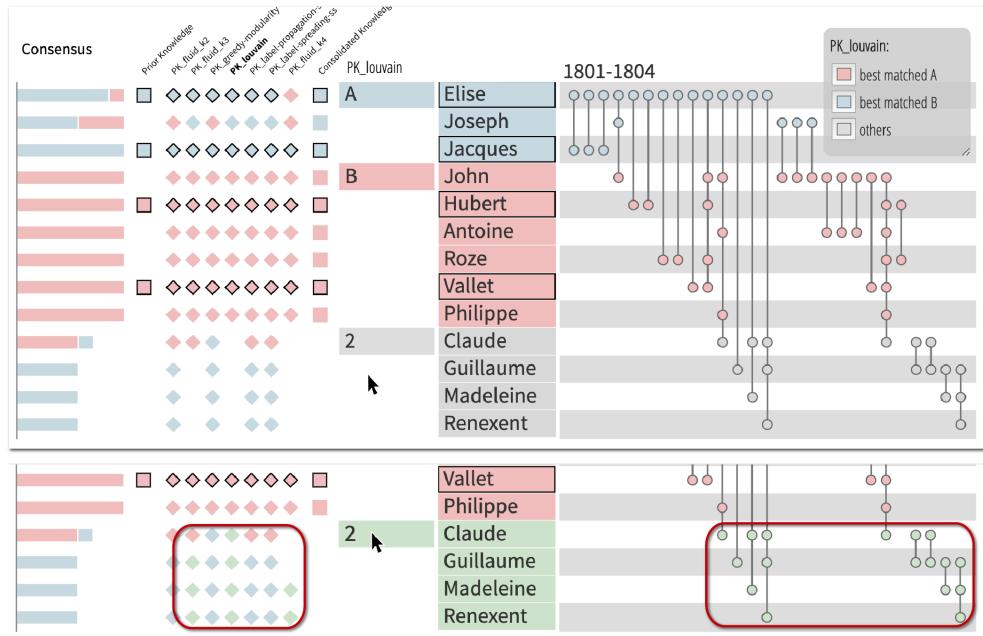


Figure 5.8 – Suggestion of extra clusters. The two PK-groups (red and blue) are validated (nodes in the consensus column are all squared). One extra clusters is proposed by the Louvain algorithm, labeled as 2. Hovering over the cluster 2, the consensus is displayed by the green diamonds. This feedback is also visible in the graph.

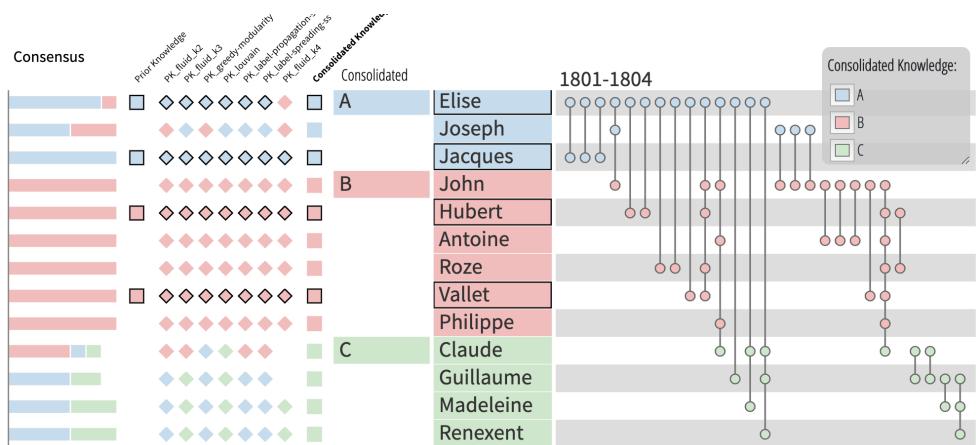


Figure 5.9 – The dataset has been fully consolidated. The persons are grouped and colored by the consolidated knowledge. The user decided to assign Claude, Guillaume, Madeleine and Renexent to cluster C, by taking into account the graph and the consensus of the algorithms.

of algorithm output and when hovering over the node. A click on the color will validate the node into the cluster having that color. If users are satisfied with the association proposed by the current algorithm, they can validate it by clicking on the cluster name. This will create a new group, so the user can classify the nodes into this new group, as seen before (subsubsection 5.3.7) : using the consensus slider, copying an algorithm result, or through manual labeling. This process is repeated for the other clusters until there are no unlabeled nodes or the user is satisfied with the partial clustering. An example of a fully consolidated dataset is shown in Figure 5.9.

5.3.8 . Wrapping up and Reporting Results

At any stage of the process, the user can finish instantaneously, either by not labeling undecided nodes, or selecting and validating the results of a single algorithm—as traditional approaches do, or by using a specified threshold of consensus and not labeling the remaining entities. The appropriateness of the choice is up to the user and should be documented in the publication.

In addition to the consolidated clustering, the output of PK-clustering consists of provenance information in the form of a table and a summary report. The table provides, for each vertex, the consolidated label, along with the labels produced by all the selected algorithms, and a description of the interaction that has led to the consolidation, such as “selected from algorithm x”, “consensus ≥ 5 ”, or “override” when manually selected by the user instead of selected from an algorithm. The summary provides counts of how many nodes were labeled using the different interactions methods and can be used in a publication. Examples are provided in the Supplemental Materials (as Fig. 2 and Fig. 5).

Clustering results can thus be reviewed in a more transparent manner, revealing the decisions taken. In contrast, traditional reporting in the Humanities rarely questions or discusses how choices were made and merely mentions the algorithm and parameters used.

5.4 . Case studies

We describe two case studies using realistic scenarios where the clustering has no ground truth solution but has consequences, scientific or practical. We also report on the feedback received from practitioners.

5.4.1 . Marie Boucher Social Network

We asked our historian colleague her prior knowledge on her network about the trades of Marie Boucher [?], composed of two main families : Antheaume and Boucher. Family ties were important for merchants, but could not scale above a certain level. Marie Boucher expanded her trade network far beyond that limit. She then had to connect to bankers, investors, and foreign traders, far outside her family and yet connected to it indirectly. As hinted in her article, Dufournaud

believes that the network can be split in three clusters : one related to the Boucher family, one to the Antheaume family, and the third to the Boucher & Antheaume company. Using standard visualization tools, she could see different connection patterns over time, but she wanted to validate her hypothesis using more formal measures and computational methods.

So she specified her hypotheses as Prior Knowledge and started the analysis. ?? (top left) shows the three PK groups : Marie Boucher for the Boucher family, Hubert Antheaume for the Antheaume family, and the Boucher & Antheaume corporation alone for the company.

After running the algorithms, 9 algorithms produced a perfect match out of the 13 executed (see ?? - left.) with the first algorithm listed an attribute based algorithm that uses the time attribute in its computation. That summary alone was found very interesting because the 3 clusters seemed very consensual among all the 9 algorithms, and furthermore, they appeared explainable by time alone..

In the PAOH view, she started by consolidating the 3 PK-groups using the amount of consensus among the algorithms as well as the graph representation and her own knowledge of the persons. At the end of this step, the Boucher, Antheaume, and Boucher & Antheaume groups were consolidated, but there were still several persons not labeled on the consolidated knowledge. She decided to review in more detail the clustering results using the *ilouvain_time* algorithm because of its reliance on the time attribute, and also because its results seemed good in the matching view. After clicking on the virtual group *Others*, the four other clusters computed by *ilouvain_time* appeared and were reviewed by hovering the mouse on the names of these new groups. She selected only one clusters she was confident about and consolidated it.

The final validated partition of the dataset is represented in ?? (right). The persons are colored and grouped by the consolidated knowledge. We can see that the final grouping makes sense in the PAOH visualization on the right. Only one person is not part of any group : Jacques Souchay. It is not unusual in historical sources to have persons mentioned without any information on them.

Our historian colleague can now publish a follow-up article validating her hypotheses. The summary report will help document where the final grouping came from, increasing trust with regard to her claims.

5.4.2 . Lineages at VAST

In the second cases study we took the role of Alice, a VAST Steering Committee (SC) member, who participates in a SC meeting to validate the Program Committee proposed by the VAST paper chairs for the next conference. One of the many problems that all conference organizers face is to balance the members of the Program Committee according to several criteria. The InfoVis Steering Committee Policies FAQ states that the composition of the Program Committee should consider explicitly how to achieve an appropriate and diverse mix [?] of : — academic lineages — research topics — job (academia, industry) — geography



Figure 5.10 – Computing the Lineages of VAST authors : Prior Knowledge from Alice and results of the clusterings matching it.

(in rough proportion to the research activity in major regions) — gender. Most of these criteria are well understood, except *academic lineage* which is not clearly defined. Alice will use the “Visualization Publications Data” (VisPubData [?]) to find-out if she can objectify this concept of lineage to check the diversity of the proposed Program Committee accordingly.

Using PK-clustering, Alice loads the VisPubData, filtered to only contain articles from the VAST conference, between 2009–2018. Only prolific authors can be members of the program committee, but highly filtering the co-authorship network would change its structure and disconnect it. Thus, she will use the unfiltered network of 1383 authors to run the algorithms and perform the matching (Step 1 of the process), even if at the end only 113 authors with more than 4 articles will be need to consolidated (Steps 2 and 3).

Alice starts the PK-clustering process by entering her prior knowledge, which is partial and based on two strategies : her knowledge of some areas of VAST, and the name of well-known researchers who have developed their own lineage. She runs the algorithms (Figure 5.10) and 5 algorithms produce a perfect match, acknowledging her knowledge of some areas of VAST. She then shows the results to other members of the SC who will help her consolidate the lineage clusters.

Her initial PK clusters are quickly consolidated, using Internet search to validate some less known authors. She then decides to create as many additional clusters and lineage groups as she can. For some authors, she decides to override the consensus of the algorithms. For example, she decides, and her colleagues agree, that Gennady and Natalia Andrienko should be in their own lineage group and not in D. Keim’s (Figure 5.11). The history of VAST in Europe, very much centered around D. Keim and the VisMaster project [?], has strongly influenced the network structure and some external knowledge is required to untangle it.

Using the *PK_louvain* algorithm as starting point, Alice creates new groups and achieves a consensus among the experts on a plausible set of lineages for VAST. She then checks with the list proposed by the program committee by entering it

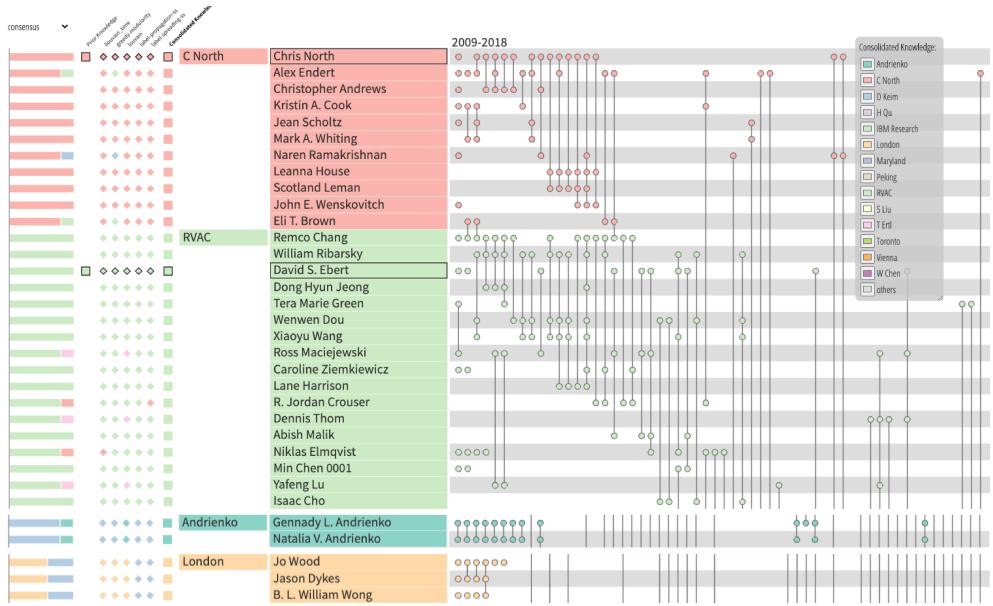


Figure 5.11 – Four consolidated groups in the VAST dataset : C North, RVAC, Andrienko and London

in on a spreadsheet with the names and affiliations. She adds the groups and their color, and sort the list by group. Alice can now report her work to the whole SC, which can check the balance of lineages according to this analysis, and decide if some lineage groups are over or under represented. By keeping the affiliations in the list, the SC can also check the balance of affiliations that is not always aligned with the lineages. The final results are available in the supplemental material of the article.

Using partitioning clustering (although with outliers) forces the algorithms or experts to make strong decisions related to lineages. But using a soft clustering (or overlapping partitions), while providing a more nuanced view of lineages, would not be as simple to interpret as coloring spreadsheet lines and sorting them ; in the end, the final selection only uses the lineage criterion among many others. Still, we believe PK-clustering can provide a partial but concrete answer to the problem of defining what the scientific lineages are.

5.4.3 . Feedback from practitioners

Although we could not conduct face to face meetings with historians and sociologists due to the COVID19 lockdown, we showed the system to three practitioners and asked their feedback through videoconferencing systems, sharing video demonstrations and sharing our screen.

They all acknowledged the pitfalls of existing systems providing clustering algorithms as black boxes with strange names and mysterious parameters. They also agreed that the current process for clustering a social network was cumbersome

when they wanted to validate the groups and compare the results of different algorithms. None of the popular and usable systems provide easy ways to compare the results of the clusterings. Usually, the analyst needs to try a few algorithms, remembering the groups that seemed good in some of the algorithms, sometimes printing the clustered networks to keep track of the different options. Still, they all confirmed that they usually stop after trying 2 to 3 algorithms because of lack of time and support from the tools. Evaluation of clusterings is long and tedious.

They were intrigued by the idea of entering the prior knowledge to the system, but acknowledged that it was easy to understand and natural for them to think in terms of well-known entities belonging to groups. They felt uneasy thinking that this prior knowledge could bias the results of the clustering and of the analysis. However, after a short discussion, they also agreed that the traditional process of picking in a more or less informed way two or three algorithms to perform a clustering was also probably priming them and adding other biases. Still, they said that they would need to explain the process clearly in their publications and that some reviewers could also stress the risks.

They all agreed that the process was clear and made sense, but they also felt it was complicated and that they would need time to master it. They said that it was more complicated than pressing a button, but that the extra work was worth it.

One historian who spends a lot of time analyzing her social networks and finding information about all the people was shocked by the idea that you could want to use an algorithm that did not match fully the prior knowledge. For us, it matters if the prior knowledge is given as constraints or preferences, but we did not want to introduce these notions in the user interface so analysts are free to interpret the prior knowledge as one or the other.

They also identified some issues with the prototype. It was not managing disconnected networks at all when we showed the demo, and they stressed the fact that real networks always have disconnected components. They were also asking about structural transformations, such as filtering by attribute or by node type. We chose not support these functions at this stage, but they can be done through other standard network systems.

They were also interested in getting explanations about the algorithms, why some would pick the right groups and others would not. Our system is not meant to provide explanations and works with black box algorithms. We wished we could help them but that would be another project. Still, when an attribute-based algorithm matches the prior knowledge, we believe that attribute-based explanations are more understandable, e.g., groups based on time, or income.

The table and summary report was added after those sessions so no feedback was gathered. We will continue to collaborate with those practitioners and help them test PK-culstering during their next social network analysis project.

5.5 . Discussion

As presented in subsection 5.2.6, the existing approaches to create clusters in social networks consider three options : standard clustering, ensemble clustering, and semi-supervised clustering. Our proposed PK-clustering approach combines aspects of the three options in order to give more control to users in the analysis loop, and allow them to have more say in the final results.

Proponents of automatic methods may argue that PK-clustering gives users too much influence on the final result as they can change the cluster assignments at will. On the other hand we know that social scientists are rarely satisfied with current clustering methods, in part because they run on graph data that rarely represent all the knowledge they have of the social network, so providing user control to correct mistakes is critical.

Traditional methods push users to believe the results of the first algorithms and parameter selection they try (typically chosen randomly). Using PK-Clustering, users can still follow blindly the results of one algorithm but PK-clustering provides a more systematic approach. It allows users to compare results, review consensus, think at each phase and reflect on decisions. Instead of passively accepting what the algorithms propose, users provide initial hypotheses—which limits the chances of being primed by an algorithm, and explicitly validate the cluster assignment of nodes, therefore performing a critical review of the automated results, yet with fast interaction to accept many suggestions at once when appropriate.

This new approach allows users to discover alternative views. For example when algorithms do not match the PK, it is an indication that the PK is being challenged and may not be correct. Users actively participate in the process of assigning, a requirement for social scientists. The report produced at the end of the analysis adds transparency by recording where the results come from for each node so decisions can be reviewed. Ultimately social scientists remain responsible for reporting and justifying their choices and interventions in their publication.

We acknowledge that bias issues are complex. The absence of ground truth limits researchers' ability to measure those biases, and no approach solves all issues yet, but we believe that PK-clustering offers a fresh perspective on those issues and will lead to results that are more useful to social scientists.

5.5.1 . Limitations

Many more clustering algorithms exist and could be added. Moreover, expanding the exploration of parameter spaces for clustering algorithms seems needed. Another limitation of the current prototype is that some algorithms do not work well with disconnected components of the graph. Unfortunately, social scientists datasets typically have many disconnected components. This issue can be mitigated by separating components into a set of connected components, run the algorithms on them, and merge the results. Our prototype runs both with node-link and PAOH representations, but it is better tuned to the PAOH representation because of its

highly readable nodes list and table format which makes the review of consensus easier. Better coordination of the table with node link diagrams and other network visualizations is needed. Further case studies will help us improve the utility of the tool as well as the provenance table and summary, which could include annotations documenting the decision process

5.5.2 . Performance

The performance of PK-clustering strongly depends on the clustering algorithms. We implemented fast algorithms to have acceptable computation times. Currently a cut-off automatically removes algorithms that have not produced a clustering after 10 seconds of computation. We ran a benchmark of the performance on the two datasets of the case studies with a laptop equipped with an Intel Core i7-8550U CPU 1.80GHz × 8 and 16 Gigabytes of memory. For the full Marie Boucher social network described in subsection 5.4.1, composed of 189 nodes and 58 hyperedges (1000 edges after the unipartite projection) it took 0.6 seconds to run all our implemented algorithms and produce the matching. For the graph of subsection 5.4.2 about the VisPubData of the VAST conference, made of 1383 nodes and 512 hyperedges (4554 edges after projection), one algorithms (the Label Propagation algorithm) took 11.37 seconds to finish and was abandoned because deemed too computationally expensive. Those two datasets are representative of the many medium size datasets historians and social scientists carefully curate (i.e., 50–500 nodes).

In order to improve the computational scalability, we will implement progressive techniques to deal with larger sizes [?]. The current user interface design for PK-clustering would allow the ranked list of algorithms to be progressively updated, and users to review a few individual algorithms first while other algorithms are still running. Of course, visual scalability is also an issue with larger datasets, as the list of people also grows. PAOHVis allows groups (like clusters) to be aggregated or expanded, so we expect that users would expand clusters one by one to review and consolidate them, while also being able to review the connections between the proposed clusters. Users can also use the automated features of PK-Clustering to consolidate the nodes (e.g., selecting one algorithm based on the ranking, or using the consensus slider to consolidate all the nodes at once). Pixel-oriented visualizations [?] would facilitate the review of consensus for a large number nodes and clusters. Classic techniques like zooming or fisheye views [?, ?] would help as long as names remain readable, which is critical to our users.

5.6 . Conclusion

In this article, we introduced a new approach, called PK-clustering, to help social scientists create meaningful clusters in social networks. It is composed of three phases : 1) users specify the prior knowledge by associating a subset of nodes to groups, 2) all algorithms are run and ranked, 3) users review and compare results

to consolidate the final clusters.

This mixed-initiative approach is more complex than a traditional clustering process where users simply press a button and get the results, but it provides social scientists with an opportunity to correct mistakes and infuse their deep knowledge of the people and their lives in the results. With simple actions such as moving a slider, or dragging over icons, users are able to interactively perform complex tasks on many nodes at once. The output of PK-clustering is—using a direct quote from a social scientist providing feedback on the prototype : “a clustering that is supported by algorithms and validated, fully or partially, by social scientists according to their prior knowledge”. Two case studies illustrated the benefits of PK-clustering.

Clustering and social network analysis remains a challenging task, typically without ground truth to formally evaluate the results. The risk of introducing bias remains always present, in this new approach as well as in traditional methods. We believe that PK-clustering offers a fresh perspective on the process of clustering social networks and gives users the opportunity to report their results in a transparent manner. The next frontier will be the analysis of dynamic social networks, that are often used in social science, and our approach will need to take into account the evolution of the communities over time.

6 - Conclusion

6.1 . Summary

In this thesis, we tried to give answers and leads to the high-level question of how VA can help historians following HSNA, in their entire process and not just focus on the analysis. For this goal, we first defined the HSNA process from data acquisition to visual analysis, to define recurring pitfalls we encountered with collaborations with social scientists. We divided the process into five steps : textual sources acquisition, digitization, annotation, network creation and network visualization/analysis, and identified recurring pitfalls for each step, such as wrongly chosen network models or named entity recognition errors. We concluded that reality, traceability and simplicity properties should be satisfied during the overall workflow as much as possible, to respectively not introduce bias and distortion in the analysis, ease the back and forth between the analysis and the processing steps and assure reproducibility of the results, and have expressive representations and tools which are simple enough to manipulate for social scientists. Specifically, we answer our question **Q1** on how to model historical documents by proposing to use which satisfy these three conditions. Leveraging this model, we first tried to find the right representations and types of interactions which could help social scientists answer their complex questions (**Q2**). For this, we developed ComBiNet using feedbacks of historians, leveraging bipartite node-link representation and maps to let social scientists explore there data modeled as , and implementing visual queries and comparisons capabilities to let them answer their potential complex questions. Finally, we proposed PK-Clusering, a new method for clustering based on social historians needs in control over algorithmic results, as a demonstration of a VA system with the right balance of usability, control and traceability. These two systems demonstrate that VA systems can help social historians in their overall workflow, and increase the traceability and control of the process while leveraging complex representations and algorithmic power.

6.2 . Discussion

We discuss in this section different limitations of our work :

Temporality. The time is a key information for historians, as they want to contextualize the phenomena they study in a period, relative to other events. This is why we encode time in our suggested model of bipartite multivariate dynamic networks, so historians can explore and analyze this dimension of their data. However, dynamic graphs are complex to visualize and analyze. In our proposed interfaces ComBiNet and PK-Clustering, time is not a central part of the interactions, and historians could therefore miss some potential interesting patterns related to it. In

ComBiNet, time is encoded as an attribute in documents nodes, and users can therefore apply filters on it, and see time distributions related to the overall network, specific documents, and filtered groups. It allows them for example to compare two periods they are interested in. However, the two layouts focus on the topology and the location first. In PK-Clustering, social scientists can build a satisfactory partition based on their prior knowledge and consensus of clustering algorithms. The prototype now consider only static clustering, which can be seen as a simplification of the real world groups which are often evolving with time. Indeed, persons often can change groups with time and clusters can sometimes merge, split, and disappear according time. Pk-Clustering is already a complex process for static clustering, but could be extended to the building of dynamic groups with the use of prior-knowledge time-dependant and dynamic graph clustering algorithms.

Put figure of dynamic layout prototype

VA for the HSNA workflow. Our key point in this thesis is to show that VA should be used in the overall HSNA workflow of historians. VA could be used to help them from data collection to their final analysis in the same environment, to ease back and forth between the steps, allowing easier exploration of different analysis goals, and better traceability/reproducibility for the overall analysis. By modeling historical documents into bipartite multivariate dynamic networks (see chapter 3), we represent the documents and their content as a network, allowing a traceability between the network entities and the original documents. If historians find errors in the network, they can rapidly trace it back from which document the errors come from, and correct it either directly in the visual interface, or in their annotation software using the unique identifier of the document. This modeling choice is a first step towards a better integration of the different steps into the same VA loop. Moreover, with ComBiNet, social scientists can apply filters to study specific visions of the network and follow multiple analysis paths on different dimensions of the data. ComBiNet therefore allow a better integration of the cleaning of the annotation, modeling and analysis/visualization steps, using the same interface. However, it does not allow complex network transformations (such as creating simple unipartite networks) nor adding new annotations in the documents texts. Historians still need to use ad-hoc methods for data collection and annotation, and may want to make other network transformations for specific analysis goals.

HSNA and Social History. HSNA is now a widely used method in quantitative history to study relational phenomena of the past, and our reflexions and tools described in this thesis aim at improving the workflow of historians following such a method. Yet, historians usually have heterogeneous and various documents when they are researching an area and era of interest, and usually apply different methods at the same time to make their historical conclusions. The core of their work consists in extracting knowledge from rigorous inspection and cross-referencing of their documents. If providing VA tools for their HSNA analysis from start to finish is useful to them, other types of analysis methods should also be implemented in their

work environments to allow them a larger set of options to make their conclusions. This includes methods like text analysis, correlation computations, and statistical testing [42].

History is also often considered a qualitative process, meaning that historians often make conclusions and hypothesis based on the reading of other sources and the qualitative analysis of their documents. VA tools which aim to encompass the whole historic workflow should be able to manage this type of analysis, for example by managing textual annotation management on the digital documents, similar to Jigsaw's feature for intelligence analysis [?]. Some quantitative methods can also let users express some of their qualitative knowledge to influence the results. For example, bayesian statistics and semi-supervised machine learning methods are based on expressing prior-knowledge which will influence the computation and results. With PK-Clustering, historians can also express their prior-knowledge and use it as a start to find meaningful clusters, by seeing how the diversity of algorithms match their vision of the data. VA tools for history should therefore let users follow both qualitative and quantitative inspection of their documents from data collection to final analysis, with combinations of several tools and prior-knowledge expression.

Diversity of Historical Documents. We elaborated our reflexions on the HSNA workflow and VA tools in collaborations with historians who base their work on semi-structured documents such as marriage acts, birth certificate, and migrations forms. These types of documents have a repetitive structure and mention people in a restricted number of relationships (spouses and witness for marriages, parents and child for birth certificate, etc.) that can be encoded as roles in a consistent manner. Historians often leverage those types of documents in their work, as they can find them in national archives. However, other types of textual documents can be used as historical sources, which can be less structured or without any predefined structure at all. One example is correspondence letters, which is a type of document often studied in history [cite](#). The content of letters is more verbose and vary from one to another, making the process of defining a set of relationships to encode more difficult. bipartite multivariate dynamic networks would therefore not necessarily be an efficient model to encode this type of data, and other network models may be a better fit. Other types of quantitative methods can also be used by historians, such as text analysis.

6.3 . Perspectives

We list in this section how this work could be extended, and interesting research directions for social history VA applications.

Dynamic Layouts and Clustering. As discussed in section 6.2,

Machine Learning, Automation, and Agency. A lot of work has been done in the recent years on machine learning, due to its rapid progress in various tasks such as questions answering, automatic driving, fraud detection, or node classification.

Machine learning has also been applied to social sciences and DH, for example for historical documents digitization [?], or link prediction [?]. Machine learning can give state-of-the-art accuracy on many of those tasks, but often set issues on the explainability and reliability of the results in real world applications. Several methods and approaches now focus on trying to explain the results of those black-box algorithms to the end-user. Similarly, research is done on how to design interactive systems which leverage machine learning algorithms to guide and advise users, who still have to take the main decisions [?]. PK-clusterering is based on this idea that machine learning should help users make decision based on automatic computations while letting them at the center of the analysis loop. ComBiNet could also be extended with machine learning features and with the same agency idea, for example to suggest social scientists recurring subgraphs in the data, that could be interesting to them. The over represented subgraphs could be a query start that the users could refine.

This idea of empowering users with the help of machine learning algorithms could be extended to the overall workflow of social historians. In their workflow, as we saw in chapter 3 they have to manually do various tasks like transcription, Named Entity Recognition, and Named Entity Disambiguation that machine learning is efficient at. VA interfaces could help social scientists do these tasks more easily by providing help and suggestions from machine learning results and interactions.

A common workflow interface. Currently, most social scientists have to use a lot of different pieces of software, files and ad-hoc processes to follow quantitative analyses. We provided two VA interfaces to help historians analyze their data and ease back and forth between the different steps of their analysis. However, historians still have to collect, annotate, and process their data manually with ad-hoc methods, and may have to convert their data to various formats when using several visual analysis softwares. All these operations make their process tedious, and usually break the traceability and reproducibility of their analyses. In the contrary, if all the processes they do is integrated in the same visual environment, it would help the flow of their analysis, increase the traceability of the results and actions, and allow them to take several explorations paths more easily. An interesting research direction would be to develop such systems, allowing social historians to collect, annotate, apply transform, analyze and visualize their data in the same environment, and with visual analytics capabilities.

6.4 . Conclusion

Bibliographie

- [1] Mobilité et conflits. Travailler sur les chantiers de construction piémontais dans la première moitié du XVIII^e 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, Sébastien 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. *Réseaux*, 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 XVI^e et XVII^e siècles. In Bernard Michon and Nicole Dufournaud, editors, *Femmes et Négoce Dans Les Ports Européens (Fin Du Moyen Age - XIX^e 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. Erdős 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 réseaux 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 Gribaudo 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] Zacharias 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.