université PARIS-SACLAY

Analyse Visuelle de Réseaux Sociaux Historiques: Traçabilité, Exploration et Analyse

Visual Analytics for Historical Social Networks: Traceability, Exploration, and Analysis

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

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

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

Thèse soutenue à Paris-Saclay, le JJ décembre 2022, par

Alexis PISTER

Composition du jury

Ulrik Brandes Professeur, ETH Zürich	Rapporteur & Examinateur
Guy Melançon Professeur, Univerité de Bordeaux	Rapporteur & Examinateur
Wendy Mackay Directrice de recherche, Univ. Paris-Saclay, CNRS, Inria, LISN	Examinatrice
Uta Hinrichs Reader, University of Edinburgh	Examinatrice
Laurent Beauguitte Chargé de recherche, CNRS	Examinateur
Jean-Daniel Fekete Directeur de recherche, Univ. Paris-Saclay, CNRS, Inria, LISN	Directeur de thèse
Chritophe Prieur Professeur, Université Gustave Eiffel	Directeur de thèse

THESE DE DOCTORAT
NNT : 2020UPASA001

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre: Analyse Visuelle pour l'Analyse de Réseaux Sociaux Historiques
Mots clés: 3 à 6 mots clefs (version en français)

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

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

Title: Visual Analytics for Historical Social Networks: Traceability, Exploration, and Analysis
Keywords: 3 à 6 mots clefs (version en anglais)

Abstract:

Historical Social Network Analysis is a method followed by social historians to model relational phenomena of the past such as kinship, political power, migrations, or business affiliations with networks using the content of historical documents. Through visualization and analytical methods, social historians are able to describe the global structure of such phenomena and explain individual behaviors through their network position. However, the inspection, encoding, and modeling process of the historical documents leading to a finalized network is complicated and often results in inconsistencies, errors, distortions, and traceability issues. For these reasons and usability issues, social historians are often not able to make thorough historical conclusions with current visualization tools. In this thesis, I aim to identify how visual analytics—the combination of data mining capabilities integrated into visual interfaces—can support social historians in their process, from the collection of their data to the answer to high-level historical questions. Towards this goal, I first formalize the workflow of historical network analysis in collaboration with social historians, from the acquisition of their sources to their final visual analysis, and propose to model historical sources into bipartite multivariate dynamic social networks with roles to satisfy traceability, simplicity, and document reality properties. This modeling allows a concrete representation of historical documents, hence letting users encode, correct, and analyze their data with the same abstraction and tools. I, therefore, propose two interactive visual interfaces to manipulate, explore, and analyze this type of data with a focus on usability for social historians. First, I present ComBiNet, which allows an interactive exploration leveraging the structure, time, localization, and attributes of the data model with the help of coordinated views, a visual query system, and comparison mechanisms. Finding specific patterns easily, social historians are able to find inconsistencies in their data and answer their questions. The second system, PK-Clustering, is a concrete proposition to increase the usability and effectiveness of clustering mechanisms in social network visual analytics systems. It consists in a mixed-initiative clustering interface that let social scientists create meaningful clusters with the help of their prior knowledge, algorithmic consensus, and exploration of the network.

Both systems have been designed with continuous feedback from social historians, and aim to increase the traceability, simplicity, and document reality of the historical social network analysis process. I conclude with discussions on the potential merging of both systems and more globally on research directions towards better integration of visual analytics systems on the whole workflow of social historians. Such systems with a focus on usability can lower the requirements for the use of quantitative methods for historians and social scientists, which has always been a controversial discussion among practitioners.

Publications

Publications related to the thesis

- A. Pister, P. Buono, J.-D. Fekete, C. Plaisant, and P. Valdivia, "Integrating Prior Knowledge in Mixed-Initiative Social Network Clustering," IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 2, pp. 1775–1785, Feb. 2021, doi: 10.1109/TVCG.2020.3030347.
- A. Pister, N. Dufournaud, P. Cristofoli, C. Prieur, J.-D. Fekete, ComBiNet: Visual Query and Comparison of Bipartite Multivariate Dynamic Social Networks, Submitted as a Fast Tracked article to Computer Graphics Forum, Wiley,
- A. Pister, N. Dufournaud, P. Cristofoli, C. Prieur, J.-D. Fekete, From Historical Documents To Social Network Visualization: Potential Pitfalls and Network Modeling. VIS4DH 2022 - 7th Workshop on Visualization for the Digital Humanities, Oct 2022, Oklahoma City, United States.
- A. Pister, C. Prieur, and J.-D. Fekete. Visual Queries on Bipartite Multivariate Dynamic Social Networks. Poster, EuroVis 2022 - 24th EG Conference on Visualization, Jun 2022, Rome, Italy. ⟨10.2312/evp.20221115⟩

Other Publications

- N. Tovanich, A. Pister, G. Richer, P. Valdivia, C. Prieur, J.-D. Fekete, and P. Isenberg. VAST 2020 Contest Challenge: GraphMatchMaker: Visual Analytics for Graph Comparison and Matching. IEEE Computer Graphics and Applications, IEEE, 2021, 42 (4), pp.89 - 102. doi: 10.1109/mcg.2021.3091955.
- S. Di Bartolomeo, A. Pister, P. Buono, C. Plaisant, C. Dunne, and J.-D. Fekete. Six Methods for Transforming Layered Hypergraphs to Apply Layered Graph Layout Algorithms. Computer Graphics Forum, Wiley, 2022, EuroVis 2022 - Conference Proceedings, 41 (3), pp.1467-8659. doi: 10.1111/cgf.14538.

Contents

1 Introduction

The goal of this thesis is to characterize and produce visual analytics tools that can support social historians conducting research on their sources—particularly when using network methods—with a focus on exploration, analysis, and traceability. Historical Social Network Analysis (HSNA) is a method—sometimes referred as a paradigm [?]—followed by social historians to study sociological phenomena through the observation of relationships of actors of the past, modeled into a network. The usage of networks as an abstraction to represent and study social relationships—such as friendships, kinship, or business ties—grew in popularity in the last 40 years [43, 138] and constitute a powerful metaphor, especially in our time when many of our digital connections and interactions use an explicit network structure[1]. This approach has first been formalized in sociology under the term Social Network Analysis (SNA) [43] and is now widely used in anthropology, geography, and history [72]. Historians leverage historical documents—which are at the core of their profession [?]—to extract relationships between actors of interest that they model with networks constructed from nodes and links that respectively represent actors (often persons) and relationships (like kinship). Using social network visualization techniques and leveraging network measures and computations, they can then test hypotheses they have and gain insight on the structural aspect of the relational phenomena they are studying [72, 149]. This approach has been followed successfully to study various subjects such as kinship [59], entrepreneurship [121], maritime routes [78], political power [105], political oppositions [104], and persecution [?]. Yet, history is considered by many as a literary and qualitative science, and many critics emerged from the history community concerning quantitative and network methods [56, 70, 80, 83], pointing to problems such as the leading to trivial conclusions, anachronisms, simplifications, and mismatches between network and historical concepts. Moreover, quantitative and network analysis are complex processes, and demand many efforts in data collection, encoding, modification, and processing before being able to make efficient observations. This thesis considers the whole workflow of social historians to better support it with visual analytics.

Social historians have to take many annotation (sometimes called encoding) and modeling decisions, concerning *what* to model from their sources into a network, and *how* to model it [23, 33], i.e., should the information of interest be represented as a node, a link, an attribute, or not reflected in the network at all, and what format should be used. Practically, they usually use ad hoc processing and analysis scripts to transform historical documents to analyzable networks, which is time-consuming, sometimes to end up with trivial or hard to interpret

[1] This analogy goes to the point that the term "Social Network" can refer both to the sociological metaphor for social relationship and the social media platforms such as Facebook.

results [2]. Still, HSNA led to many highly regarded studies with thorough conclusions, such as the study of families of power in Florence by Padgett and Ansell where they explained the rise of the Medici family through its central position in the economical, political, and trading networks of powerful families [105] or Gribaudi and Blum work on the social and professional shift during the 19th century in France [55].

The usage of visualization to graphically display networks is common in SNA[2] as it allows to unfold the structure of networks to the eyes, thus letting social scientists confirm hypotheses they had when collecting and exploring their data as well as gaining new insight through the discovery of interesting patterns and trends [24]. Images of networks also constitute an efficient mean of communication, especially in scientific productions [42]. Many visualization techniques and softwares have thus been developed since the birth of SNA, but most popular tools are usually not designed for historians specifically, meaning that they do not regard on the provenance and process leading to the network, and focus on analysis aspects only. Moreover, they usually enforce simple network models without proposing exploration mechanisms, beyond allowing to look at the network structure and computed measures. In result, many HSNA studies show a plot of their network and describe it qualitatively, often by identifying the central actors—sometimes with the help of centrality—but do not go beyond that [81]. *In this thesis, I therefore investigate how visualization can support social historians in their work, first during the pre-analysis process and secondly during the analysis step, with the right level of expressiveness, usability, and traceability.*

1.1 Social History and Historical Social Network Analysis

Social History has evolved in the 20th century with a focus on socio-economical questions and a methodological usage of sources. Quantitative and network approaches have respectively been used since the 1960s and 1980s with the development of computer science, and expanded the methods of social history. Yet, social historians are still challenged to use computer-supported tools, partly because they do not fit their research process.

HSNA is now widely used by historians to study relational phenomena like kinship, business, and institutions of the past. 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 [10, 113]. Beforehand, History was mainly political and event-centered, as the majority of

[2] Historians and sociologists following network analyses typically use similar techniques and tools for analyzing their data. The difference between SNA and HSNA hence come from the provenance and process leading to the construction of the network. I therefore use the SNA acronym for practices common in both fields and the HSNA acronym for history specificities.

work 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 [113]. Social History shifted the focus by aiming to link together sociological, economical, and political issues and by placing individuals at the center of these questions [?]. 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 such as demographics [63] and economics [52]. 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, especially in sociology and anthropology [?]. A network is an abstraction based on graph theory concepts which can be used to model phenomena based on relationships (called links) between entities (called nodes).

Sociologists appropriated this concept to model social relationships between agents of interest, allowing them to study the sociological structure of groups of interest—such as families, institutions, and companies—and concepts like friendship, oppression, and diffusion using real world observation and mathematical computations. This SNA approach allows analysts to ground results in formal network measures and metrics based on real observations instead of relying on traditional social categories such as age, job, and gender [43]. This shift in the object of study from traditional social classes and aggregates to the observation of relationships of individuals remind the microhistory movement [49] which theorized that following the life of single individuals and small groups enable the making of higher level conclusions about the social structures they live in. Social historians followed this tradition and started to appropriate network concepts to study relational aspects of the past and formalized it under the term Historical Network Research or Historical Social Network Analysis [149]. However, historians do not have the possibility to run surveys or directly observe interactions of the past and are thus constrained by the information contained in historical documents they find in archives. These documents can be anything mentioning social relationships between actors of interest, such as marriage acts, birth certificates, census, migration acts, business transactions, journals. After selecting a corpus of documents, they typically read and inspect in depth several documents while taking notes to have a deeper insight on the content of the sources, which allow them to start eliciting hypotheses. Following this exploration phase, they manually annotate each document and encode the desired information—the mention of persons and their social relationship in the case of a network analysis. This is a long and tedious process that can result in small to large networks that they analyze using network measures to make conclusions on the structure of social groups or social behaviour of individual of interests.

The investigation and reading of the historical documents is therefore an ex-

ploratory process, where historians start to generate sociological hypotheses from the continuous extraction of insight and revelations of this process, similarly to grounded theory [50]. Once they finalised a network, they can test their hypotheses using qualitative or quantitative methods—based on statistical and network measures. Lemercier and Zalc write "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." [81] 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, in a way similar to the competing hypotheses workflow of Intelligence Analysis [30]. By pointing to evidence supporting or refuting hypotheses, they can give insight into the level of the plausibility of different claims.

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 data [19]. Using visual abstractions (such as size, color, and position) to display abstract data allows us to rapidly see structure and patterns otherwise hidden in raw text and numbers. As data keeps growing in size with time due to the increase of hardware and storage capabilities, visualization is a powerful tool to gain insight into the underlying structure of various complex datasets.

Visualization has traditionally been used for confirmatory and communication purposes, particularly in empirical sciences [**?**]. By showing data visually, analysts are able to confirm or refute hypotheses and communicate their findings in scientific productions.

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 in the 1960s [141], 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. Social network visualization is used for communication of findings in the field, but is also often following this exploration process as showing the network visually allows social scientists to reveal the structure of their data. As freeman writes "Images of social networks have provided investigators with new insights about network structures and have helped them to communicate those insights to others" [42]. Social scientists very often represent their data using node-link diagrams, that we find in every production of reference in the field [79, 138, 148].

Figure 1.1 shows a node-link representation of the network constructed by Padgett and Ancell in their work on the Medici. At that time, diagrams were often drawn by hand, practice which have now been replaced by automatic layout

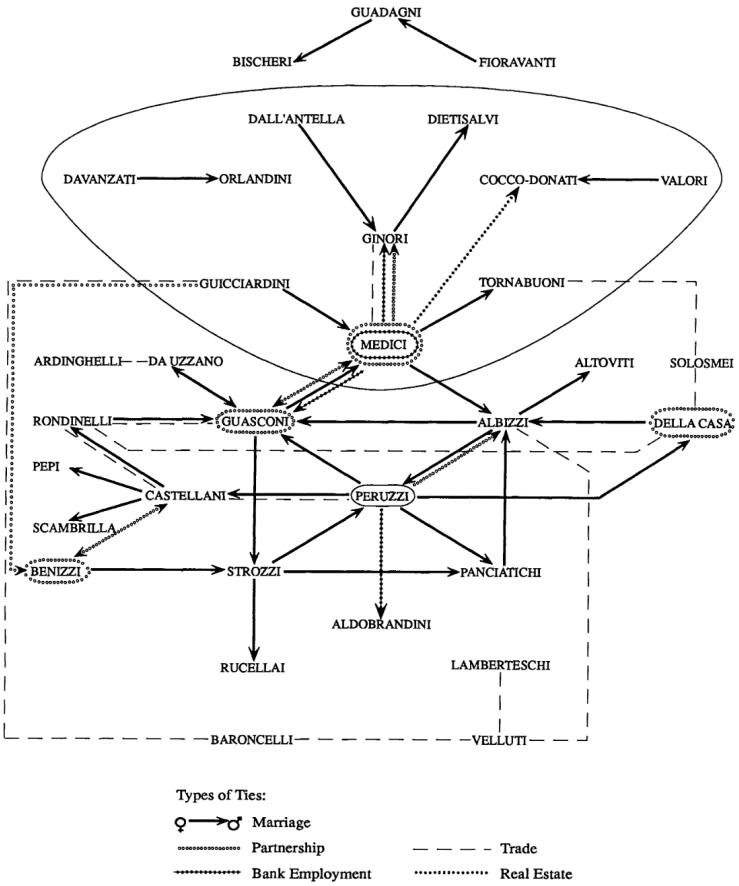


Figure 1.1 – Marriage, partnership, trading, banking, and real estate networks of the powerful families of Florence from [105]. We can see the central position in the network of the Medici Family.

algorithms.

Most visual software for SNA such as Gephi [5], Pajek [106], NodeXl [133], or Ucinet [?] are based on this representation, and allow an exploration of the data with the help of basic interaction mechanisms and the computation of network measures.

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 interactive analysis workflow to create knowledge [71].

Figure 1.2 illustrate the schematic process of VA: the coupling of visualization and data mining models operated by the user through interaction lead to the generation of knowledge "extracted" from the data.

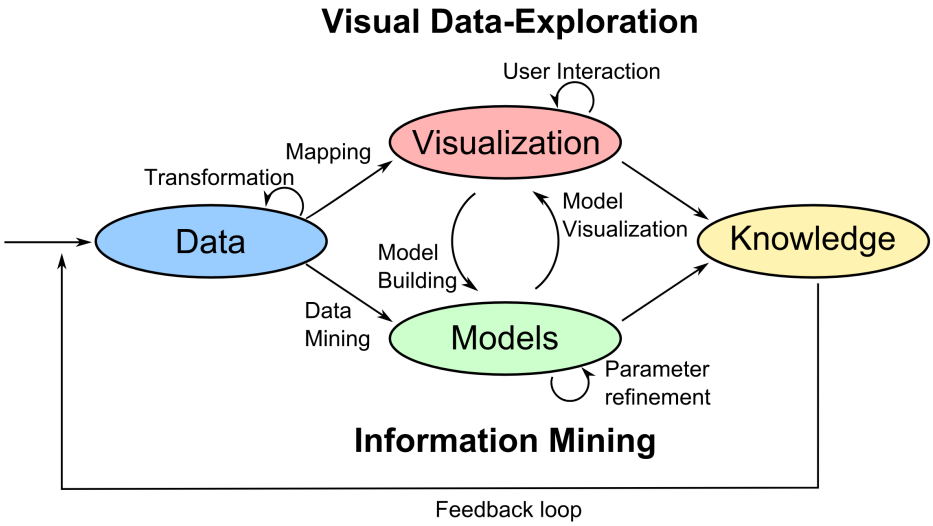


Figure 1.2 – Abstraction of the VA process. It is characterized by continuous interactions between the data, visualizations, models, and knowledge. Image from [71].

If most widely used visual interface for HSNA do not yet provide complex interactions or data mining capabilities, more recent tools are oriented towards VA, as complex interactions and data mining capabilities can be a powerful support for social scientists to gain insight on the structure of their network, especially that the data they study keep growing in size and complexity [69].

1.3 Historical Social Networks Visual Analytics

Most visual tools for SNA are designed for the analysis of already curated networks, without taking into account the context in which those networks have been produced, where they come from, and the workflow that led to their creation. Currently, social historians spend a lot of time in their data acquisition, processing, and encoding steps which lead them to the construction of a network [82]. They typically visualize and analyze their network at the end of this process, first to verify hypotheses they formulated during the inspection of their sources, then to gain a better view of the structure of the network, allowing them to potentially generate new hypotheses [34]. 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 [2, 80]. Social scientists usually visualize their network using SNA tools like Gephi, 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 [2]. They, therefore, have to go

back and forth between the visualization software and the encoding 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 that 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 visualization of a finalised network. VA could also help social scientist reflect on their network modeling process as historical documents can be modeled in various ways [23]. Yet, most visualization tools used for HSNA enforce simple network models, which visualization and analysis can often lead to trivial or distorted conclusions [80]. Furthermore, the data mining capabilities proposed in those softwares are often based on black-box algorithms, which can be hard to interpret for social analysts. For example, clustering algorithms are often included in such systems, letting social scientists partition networks into groups, but many algorithms exist in the literature, potentially giving diverse results. Scientists often run several algorithms until finding a satisfying enough partition, which can bias the result of an analysis[3]. Usability and traceability of the results are therefore primordial in VA interfaces aimed at supporting social historians in their analysis.

1.4 Contributions and Research Statement

The goal of this thesis is to characterize how VA can support social historians in their HSNA process and present proofs of concepts of tools supporting it. Most social network visualization tools are agnostic to the process of social historians leading to a polished network, even though it has an high impact of the network model and structure. Using visualization only at the end of the process often reveals potential errors, inconsistencies, or mismatches between the network model and analysis goals [2]. Moreover, due to lack of usability and interaction mechanisms, social historians often simply visualize statically their network and partially describe their structure, leading to conclusion which would have been easier to reach with simpler methods [39].

VA could therefore 1) assist social historians in their overall workflow, starting at the documents' acquisition to the final analysis step, with the help of data mining and interaction mechanisms in the data acquisition, encoding, modeling steps, and 2) provide exploration and analysis mechanisms to answer complex historical questions, beyond simply plotting the network with a node-link diagram.

The goal of this thesis is hence to give answers to the high-level question "How can VA support social historians in their entire HSNA process?". To answer this question, I first characterize the HSNA process from start to finish from discussions and collaborations with social historians, with the goal of identifying pitfalls that

[3]We did not find any scientific production describing these potential issues when following such analyses. However, many social scientists refereed to these problems with us in informal discussions.

regularly arise and characterizing social historians' needs. From this, I give answers and directions—illustrated by proof-of-concepts—to three questions concerning the modeling aspect of HSNA and how VA and automatic tools can support social historians in different parts of their process:

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

Q2: What representations and interactions would allow social historians answer complex historical questions—with a focus on usability ?

Q3: How to design VA tools and interactions that leverage algorithmic power but keep historians in control of their analyses and biases?

In chapter 3, I start by describing the HSNA workflow and identify recurring pitfalls we encountered in our collaborations with historians and answer **Q1** by proposing a network model for modeling historical documents. In the following chapter 4, I 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 find errors easily, transform their network data, answer their questions, and generate interesting hypotheses. Finally, in chapter 5, I propose PK-Clustering, a mixed-initiative clustering technique for social scientists based on their prior knowledge, algorithmic consensus, and traceability of results, as a concrete example of a system providing answers to **Q3**.

2 Related Work

Social historians rely on textual historical documents to study social groups through their structures and socio-economic characteristics in societies of the past [?]. They read and analyze documents they can find from a period and subject of interest, and make their conclusions through deep inspection and cross-referencing of the information they found. Several methods have been developed in History to extract and analyze the information contained in the documents in a methodical way [?], based on qualitative or quantitative methods—among which HSNA is now widely popular [?]. HSNA is a method consisting in modeling the relational information mentioned in the documents—such as family, business, or friendship ties—in a network, to be able to characterize and explain social behaviors through the description of the network's structure [72, 149]. This approach is directly inspired by SNA, which is a well-known method that sociologist theorized to understand and describe real world social relationships modeled as networks [43, 127]. Historians appropriated this method, by extracting relationships from historical documents. The specifics of HSNA in contrast of its sociology counterpart is therefore the modeling of the network from the historical documents—which are at the core of the historical work [113]—and the integration of the temporal dimension which is often disregarded in traditional SNA but central in history. Once they successfully constructed a network—which is a long and tedious process—they typically use network measures and visualization techniques to confirm or generate new hypotheses [80]. Visualization let them unfold the structure of their data, revealing potentially interesting social patterns between actors of the network. Analytics and visualization systems for SNA typically allow exploration of such data with the help of interaction, network measures, and data mining capabilities such as clustering directly implemented in the interfaces. Yet, most HSNA studies only give qualitative description of their network—which Rollinger call "soft" or "informal" network research—probably due to usability and formalism issues [2] . The coupling of visualization and data mining through interaction to support the generation of knowledge has been described as VA and can therefore provide support to social historians for their network construction, but also to go beyond simple qualitative description of their data. In this chapter, I first present a general overview of the field of visualization in §2.1 to share its utility and potential for social history. Then, I present the social history discipline with its use of quantitative methods in §2.2, before describing in depth how network analysis has been applied in the field in §2.3. Finally, I present in §2.4 how visualization and VA have been used in the context of HSNA, along the most popular systems currently used by social scientists and their limitations.

2.1 Visualization

Visualization is often defined as "the use of computer-supported, interactive, visual representations of data to amplify cognition" [19]. Graphically displaying data allows us to leverage our visual system to gain a better acquisition of knowledge, leading to better decision-making, communication, and potential discoveries. The field of visualization can be split in three sub domains: **Scientific visualization** focus on visualizing continuous physically based data such as weather, astrophysics, and anatomical data, sometimes produced with simulations whereas **Information Visualization** is centered around the visualization of discrete abstract data points, often multidimensional. **Visual Analytics** emerged later from Information Visualization by mixing data mining and more complex analysis process with traditional information visualization displays. I focus in this thesis on the two former branches of visualization, as social scientists use both information visualization and VA systems to gain insight on the structure of the networks they are studying.

2.1.1 Information Visualization

Information Visualization focus on displaying abstract data to amplify cognition and gain insight on real world phenomena [19]. History is filled with classical examples of visual data displays which helped understand better specific events, such as Minard's map of Napoleon's march in Russia [45], or Snow's dot map of cholera cases in London which showed the proximity between street pumps and cholera infections [135]. If several examples of information visualization can be found thorough history, it mainly developed as a scientific field in the 1960s with Tukey's work on data analysis and visualization [140] and Bertin's publication of Semiology of graphics [8].

In this foundational work, Bertin described and organized the different visual elements usable in graphical information displays, and linked them to data features and relations types. An illustration of this work of categorization for network data is illustrated in Figure 2.1. Michael Friendly writes that "To some, this appeared to do for graphics what Mendeleev had done for the organization of the chemical elements" [46]. The development of computer science and the rise of hardware capabilities during the same time created a big need for data visualization. The amount of data stored increased exponentially [?] and descriptive statistics were not enough to understand the underlying structure of the amount and diversity of produced data. Visualization, leveraging the human visual system, allowed to rapidly see the hidden structure of a dataset and detect interesting and unexpected patterns very often unseen with classical statistical methods. One classical illustration of this is Anscombe's quartet [4] which consists of four datasets of 11 points in $\mathbb{R}^2$ with the same statistical measures (mean, variance, correlation coefficient, etc.) but with very different structures, that plotting the data show immediately. The four datasets are illustrated in Figure 2.2.

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

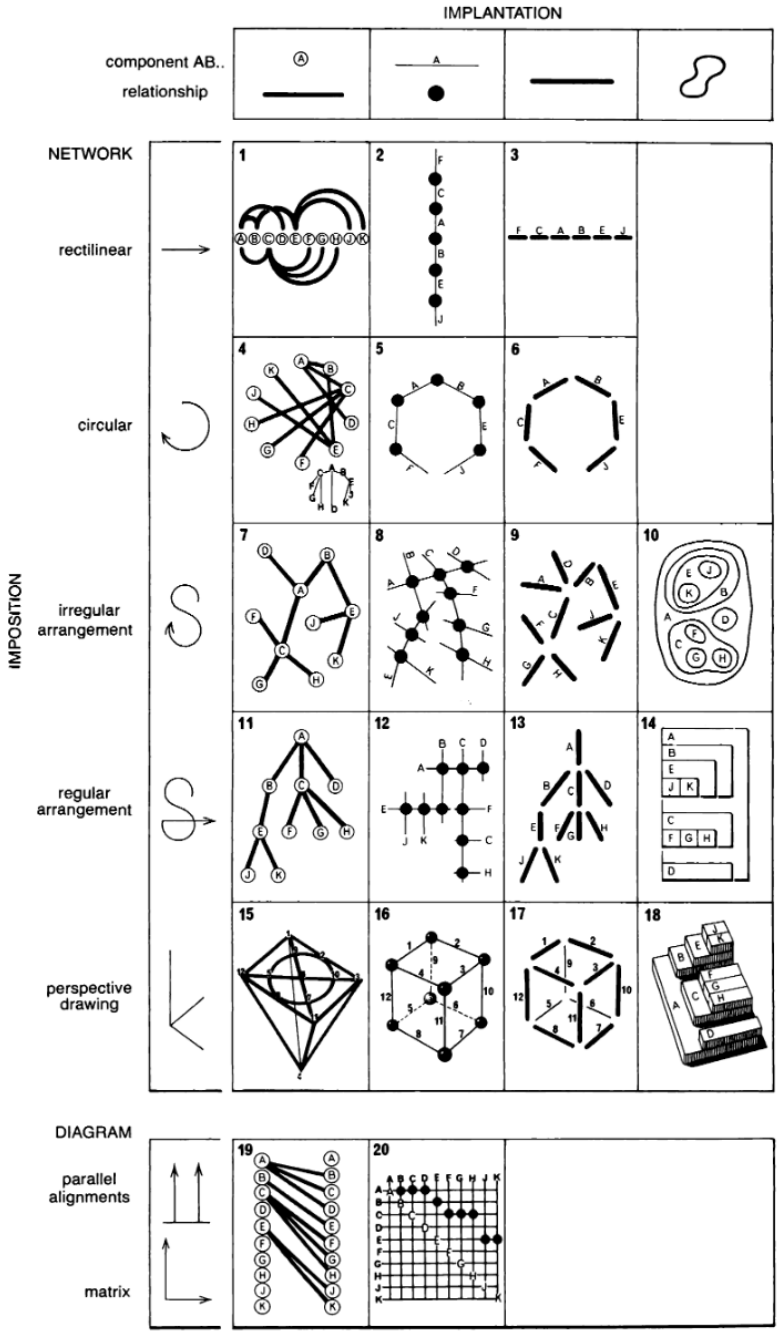


Figure 2.1 – Categorization of visual variables which can be used to represent network data, resulting in many different network representations. Image from [8].

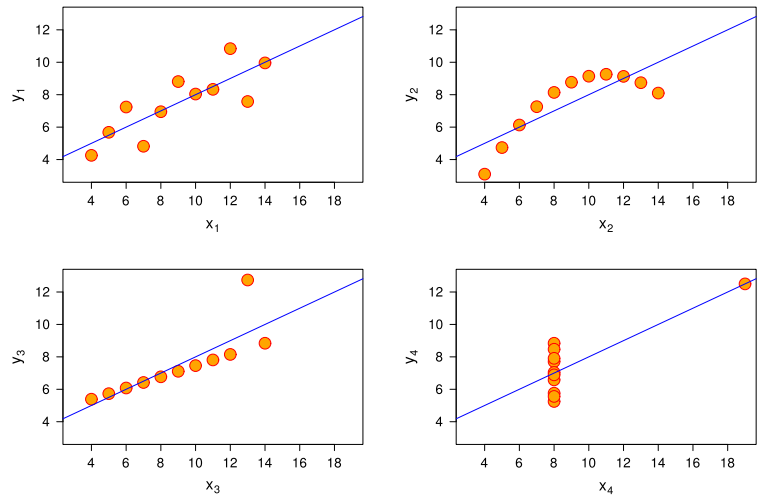


Figure 2.2 – Anscombe quartet. The four datasets have the same descriptive statistics (average, variance, correlation coefficient) but very different structures. Image from [4].

versity of data produced, such as multidimensional, temporal, spatial, or network data [?]. Instead of using taxonomies classifying graphics into categories such as histograms, pie charts, and stream graphs, some theorized how to describe graphics in a more systematic and structural way. In 1993, Wilkinson extended Bertin's work and developed the Grammar of Graphics [?] as a way to describe the deep structure unifying every possible graphics, thus allowing to characterize and create graphics using common terms and rules. In this framework, a graphic can be defined as a function of six components: data (a set of data points and attributes from a dataset), transformations (statistical operations which modify the original data, e.g., mean and rank transformations), scales (e.g., linear and log scales), coordinate systems (e.g., cartesian and polar coordinate systems), elements (graphical marks such as rectangular or circular marks, and their aesthetics, e.g., color and size), and guides (additional information such as axes and legend). Many well-known visualization toolkits are now based on this framework, such as vega [125] and ggplot [?], as it allows greater expressiveness and reusability for graphic creation. Visualization allows to gain insight on the structure of a given data, and has traditionally been used for confirmation and communication purposes, for example to verify hypothesis on empirical sciences, and later on to communicate findings, first to scientific peers, and nowadays to broader audiences for example through the means of data journalism [?].

2.1.2 Visual Analytics

VA consist in the coupling of visualization and data mining techniques to better support user in their knowledge generation process through continuous interaction

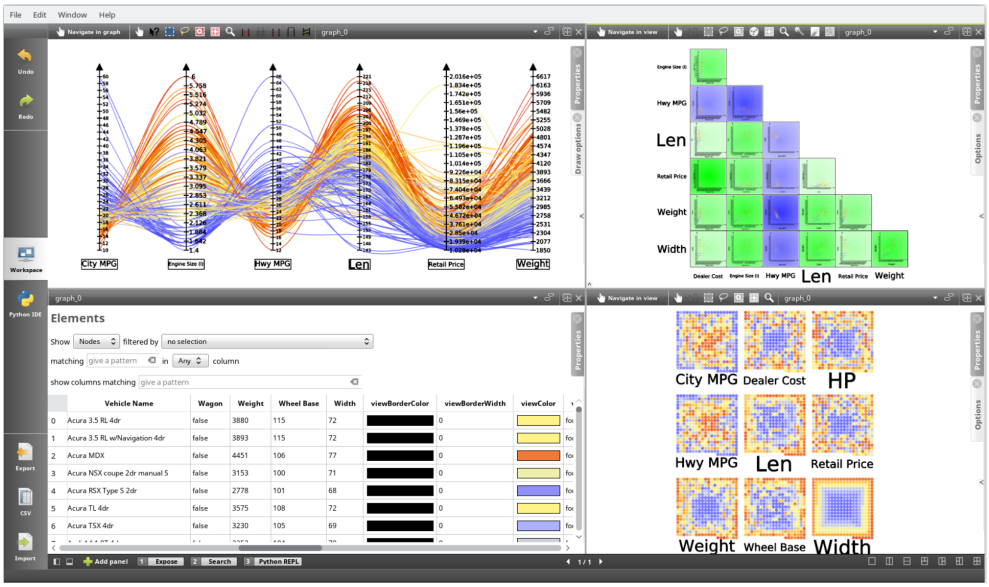


Figure 2.3 – TULIP software designed for application-independant network visual analytics [**?**]. The view shows a dataset among multiple interactive coordinated views. Users can also apply data mining algorithms on the data to extract interesting patterns.

with the data and statistical models [?]. It draws inspiration from exploratory visualization, interaction, and data mining. The process of exploratory visualization to gain new insights on the general structure of the data and potentially generate new hypotheses has been characterized by Tukey in 1960 as *Exploratory Data Analysis* [141]. It consists in trying to characterize the structure of a dataset with the help of continuous visualization and statistical measurements of different dimensions of the data. Visual exploration is enhanced by direct manipulation interfaces through interaction and usually follows the information-seeking mantra formalized by Schneiderman: "Overview first, zoom and filter, then details-on-demand"" [?]. It allows users to first have a visual overview of the data and get an idea of its overall structure, to then change the point of focus to highlight interesting patterns with the help of filtering, querying, sorting, and zooming mechanisms. As the average size of datasets keeps growing, exploratory tools are often needed to make sense of large datasets and generate pertinent hypotheses.

More recent visual exploration interfaces also incorporate automatic analytical tools along with graphical displays, letting users apply data mining algorithms directly in the exploratory loop. This coupling of visualization and analytical methods such as data mining has been defined as Visual Analytics (VA) and is still a very active research field. Keim and al. define it as "a combination of automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data" [71].

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

2.2 Quantitative Social History

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

2.2.1 History, Social History, and Methodology

The concept of History is hard to define as its practice and codes highly evolved through time. Prost writes that "History is what historians do. The discipline called history is not an eternal essence, a Platonic idea. It is a reality that is itself historical, i.e. situated in time and space, carried out by men who call themselves historians and are recognized as such, received as history by various publics [113]." Retrospectively, History of a given time can thus be characterized by the different historical work produced at that time. Nevertheless, history can be characterized as the collection and study of historical documents to study and describe the past. As Langlois and Seignobos write, "The search for and the collection of documents is thus a part, logically the first and most important part, of the historian's craft" []. History emerged as a field with its own rules, conventions, and journals in the 1880s from faculties of letters, to counterbalance previous history works which were judged as too "literary" [**?**]. A that time and until now, two facets characterize the field, which are sometimes overlapping: one is political whereas the other one is methodological. The former aspect of history serves to create a shared story for the studied country and a sense of unity to its citizens. Antoine Prost says that "it's through history than France thinks itself" [113]. The latter aspect of history constitute a methodology to describe the past through methodical inspection of historical sources, in the aim of inferring dated facts about the past and trying

to minimizing possible bias. Historical documents are thus at the core of the work of historians and having to cite historical documents and previous peers work to new claims is primordial to be considered as rigorous History work. However, methodological and epistemological facets (how historians should read and analyze their sources, how to cite them, what to report/not report, and what is the status a proof) of History have not been studied and discussed for a long time, until the end of the 1980s. Some historians were interested in historiography [18], but none were going to philosophical and epistemological reflexions of the History discipline. For Lucien Fèbvre, philosophising was even constituting a "capital crime" [40, 113].

Retrospectively, we can still observe shifts in the objects of study of historians through time, and their relation to sources. History was at first mainly event-centered and was focusing in characterizing central figures of the past like rulers and artists or shed light on central events like wars or political crisis. This narrative approach to history has been criticized for its open interpretation of historical documents, which can introduce bias from the authors [14]. In the 1930s, March Bloch and Lucien Fèbvre detached from traditional history by creating the "Annales school" (École des Annales) which aimed at placing the human as a component of a broader sociological, political, and economic system with influences between each other [16]. They strongly advised to exhaustively search from archives, to ground historical results in documents, texts and numbers. This new way of studying past events and societies became successful in a profession in crisis, by bringing a new lens of study on various societal subjects more grounded in sources and with a better intelligibility. This school of thought can be seen as one of the biggest milestones for Social History, which focuses on the socio-economical aspects of societies and their changes through time, rather than an event-centric view of History. For example, in his thesis, Ernest Labrousse—a well known figure of Social History—tries to describe and explain the economic crisis of France at the end of the "Ancien Régime"[1] through the evolution of the economic power of different social groups such as farmers, workers, property owners etcinstead of solely describing memorable facts about the period [76]. Social History continued to evolve since the 1930s, introducing new methods and concepts, but always with the goals to describe periods and historical facts through a sociological lens and with a strong focus on sources and traceability.

2.2.2 Quantitative History

With the development of statistical methods and Computer Science, quantitative approaches of History emerged in the 1960s with the goal of analyzing numeric data directly extracted from historical documents. Economists led this first wave of quantification by studying past events using economical concepts and data. This approach, called "new economic history" or "cliometrics" was popu-

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

larized by Fogel's study on the economic impact of the development of railroads in America [?] and Fogel and Engerman's controversial work on the economy of slavery [?]. In the latter study, they extracted numbers of a sample of 5000 bills of slave sales from New Orleans to support the controversial claims that slavery was economically viable and that slaves had a decent material life, which brought up heated debate among the scientific community and the broad audience [?]. These kinds of approaches rapidly started to be used in other related domains such as demography, social history, and political history, sometimes rebranded as "new social history" and "new political history" [82] As extracting the data from raw documents and uploading it in computers—which were shared among whole departments—was very time-consuming at that time, "new history" projects often relied on a high division of labor among researchers, assistants, and students who operated with punch card operators [?]. Many saw the future of social sciences in computer programming, as Le Roy-Ladurie who wrote in 1968 "The historian of tomorrow will be a programmer, or he will not exist" [81].

However, quantitative methods started to be criticized in the 1980s with a vague of disillusionment, for several reasons. Stone was the first to raise his voice in 1979, after participating himself in several of those ambitious projects: "It is just those projects that have been the most lavishly funded, the most ambitious in the assembly of vast quantities of data by armies of paid researchers, the most scientifically processed by the very latest in computer technology, the most mathematically sophisticated in presentation, which have so far turned out to be the most disappointing" [?]. First, many researchers of this first wave dispensed themselves of source criticism, leading to simplification, anachronisms—such as using modern analytical categories and indices like the GDP—, and taking the numeric data from historical documents as objective. These problems could be in part explained by the fact that the work process was highly divided, meaning that the people analyzing the data did not necessarily inspect and read the original historical documents in depth. Secondly, the popularity of these methods made practitioners forget about the many biases inherent to statistics, such as the sampling bias, or the fact that historical data is essentially uncomplete data. This resulted in the computation of long data series and aggregates which were sometimes nonsensical given the gaps in the sources [81]. Finally, many historians raised their voice against the study of long-term trends instead of focusing on specific events and individuals. They challenged aggregations procedures and its assumptions, trying to go back to a more complex history by pointing that phenomena have to be studied and understood through several scales [?]. Indeed, computing correlations and aggregates at a national level greatly simplify complex phenomena, and misses specific group and individual related behaviours. Still, if their adoption remains slow and sometimes criticized among historians, quantitative methods provide tools to store, explore, and analyze historical documents systematically if used appropriately (i.e. not trying to bias the analysis, and not losing the trace of the original sources),

especially that those methods highly evolved since the 1960s.

2.2.3 Digital Humanities

Digital Humanities is sometimes described as the second wave of computational social sciences [81]. The term has gained popularity since the 2010s and refer to "*research and teaching taking place at the intersection of digital technologies and humanities. Digital Humanities aims to produce and use applications and models that make possible new kinds of teaching and research, both in the humanities and in computer science (and its allied technologies). Digital Humanities also studies the impact of these techniques on cultural heritage, memory institutions, libraries, archives and digital culture.*" [**?**]. If the first wave of computational social sciences focused a lot on statistical methods such as regression models, correlation testing, and descriptive measures (mean, median, and variance) to make conclusions, digital humanities focuses more on the use of digital tools for exploration, teaching, and communication of humanities datasets and concepts, leveraging design, infographics, and interactive systems [**?**]. In the context of historical research, the term Digital History have been coined as "*an approach to examining and representing the past that works with the new communication technologies of the computer, the Internet network, and software systems. On one level, digital history is an open arena of scholarly production and communication, encompassing the development of new course materials and scholarly data collections. On another, it is a methodological approach framed by the hypertextual power of these technologies to make, define, query, and annotate associations in the human record of the past. To do digital history, then, is to create a framework, an ontology, through the technology for people to experience, read, and follow an argument about a historical problem.*" [**?**] Research which label itself as Digital History pivot around the curation and digitization of historical archives, the identification of historical concepts through computational and exploration methods, but also their communication to the general audience through digital technologies.

Many Digital History projects are thus multidisciplinary by essence and involve several teams of researchers, such as the Republic of Letters project which consisted in digitizing, storing, and exploring letters of scholars across the world, in a common hub and using shared visualization tools [**?**]. It resulted in the elaboration of curated datasets and visualizations concerning the correspondence of various scholars such as Voltaire, Benjamin Franklin (see Figure 2.4), or John Locke, accessible in the same place by researchers and the general audience. With modern technologies and infrastructures, it also becomes possible to study large historical databases—often labeled under the term "big data"—as with the Venice Time Machine project [69] which aims at digitizing and analyzing thousands of documents from the archives of Venice to understand the political, geographical, and sociological dynamics of the cities across generations and centuries. Yet, some historians raised concern about this type of project, fearing that it could rapidly bring the same type of issues that we saw during the first wave of quantification, especially for big projects involving

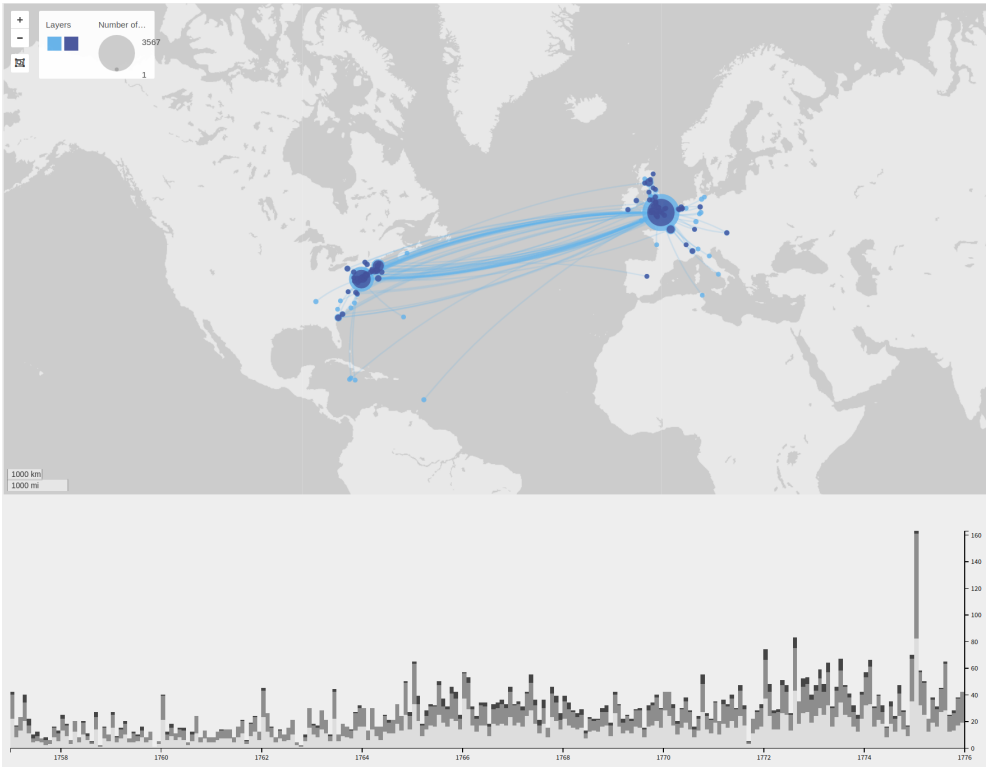


Figure 2.4 – Correspondence letters of Benjamin Franklin and his close relationships, using a map and an histogram, accessible online on the republic of letter website [**?**].

many actors and high ambitious goals [81].

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

2.3 Historical Social Network Analysis

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

2.3.1 Sociometry to SNA

One of Sociology's main goals is to study social relationships between individuals and find recurrent patterns and structures allowing to generalize on how social relations operate, and what are the social specificity of specific groups and individuals [127]. Traditional methods try to answer those questions using classical social classifications such as age, social status, profession, and gender, typically collected from surveys and interviews. Criticism pointed that this type of division is often partially biased and comes from predefined categories which are not always grounded in reality [43], and that using random sampling of individual with such methods remove them from their context. The sociologist Allen Barton wrote in this regard "For the last thirty years, empirical social research has been dominated by the sample survey. But as usually practiced, using random sampling of individuals, the survey is a sociological meatgrinder, tearing the individual from his

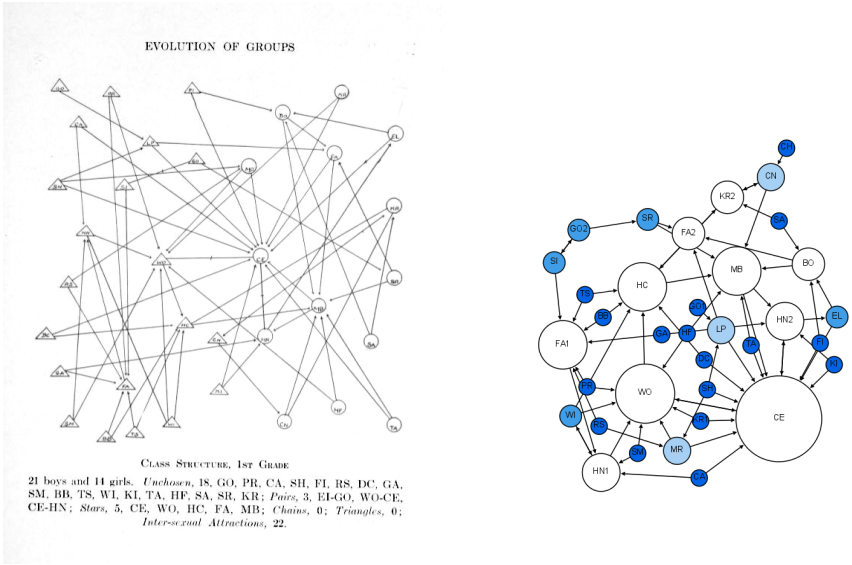


Figure 2.5 – Moreno's original sociogram of a class of first grades from [94] (left). The diagram shows 21 boys (triangles) and 14 girls (circles). The same sociogram using modern practices generated from Gephi from [53] (right). The color encodes the number of connections incoming.

social context and guaranteeing that nobody in the study interacts with anyone else in it" [?]. Sociometry is considered one of the bases of SNA and had the goal of redefining social categories through the lens of real social interactions and ties between persons, that sociologists wanted to observe in real conditions. It is in the 1930s that Moreno started to develop this new method by trying to depict real social interactions as a way to understand how groups and organizations were socially structured [95]. He developed sociograms to visually show friendships between people with the help of circles representing persons and lines modeling friendships.

Figure 2.5 shows one of Moreno's original sociograms to depict friendships in a class of first grades (left). Sociometry tremendously helped disseminate the metaphor of networks to model and understand social structures and phenomena. It was during the 1960s that sociologists and anthropologists took these concepts further and formalized SNA using graphs[2] and mathematical methods [?, 43], following the emergence of Graph Theory studies in the 1950s by Mathematicians such as Erdős [37]. Sociologists already had structural theories of social phenom-

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

ena, and they rapidly saw the potential of graphs to model social relationships between actors of interest. Typically, a graph is noted

$$G = (V, E) \tag{2.1}$$

with V a set of vertices representing the actors of interest—typically persons—and $E \subseteq V^2$ a set of edges modeling social relationships. This simple model which do not take into account the diversity and extent of social relationships still allows the characterization of the sociological structure of groups and institutions—which is the primarily focus of SNA [43, 127]. More complex network models have been proposed with time to better take into account concrete properties of social relationships. We discuss those more in depth in §2.3.4.

Graph theory brought a panoply of concepts and methods to characterize the structure of networks, that sociologists such as Coleman started to codify to use in a sociology setting [21]. The use of network measures let sociologists explain social phenomena through the formal description of real observations of relationships modeled as network.

2.3.2 Methods and Measures

Many measures and algorithms have been proposed in the network science and SNA literatures to characterize the structure of simple networks as defined in Equation 2.1 and relate it to social behaviours and phenomena [127, 138]. Network measures are either global or local, which allow to either make high level conclusions on the general structure of social relationships or individual behaviours. Widely used global measures include for example the density and the diameter, which give insight on the sparsity of the network and how distant on average are two random pairs of nodes. Conversely, local measures give information on the structural position of a node compared to the rest of the network. Centrality—probably the most used local measure—allows to formally compute a measure of how important or central are individuals in the network [**?**]. As defining what an important node is ambiguous, several types of centrality have been proposed such as the degree, betweenness, and closeness centrality, which respectively measure the number of connexions, how nodes connects different groups, and how close are the nodes compared to the rest of the network.

More generally, sociologists aim at identifying recurring patterns of sociability between actors, and link it to other behaviours, measures, or qualitative knowledge. These patterns can for example be small unconnected components, cliques, or bow-ties structures. Groups of nodes similarly located (central or distant) and having similar shapes are sometimes referred as structurally equivalent [80]. Instead of observing complex shapes, network scientists have also been interested at studying relationships at the lowest possible scale, i.e., observing relations between sets of 2 and 3 nodes at one, also called dyads and triads [148]. This reflects on Simmel's formal sociology, where he already referred to dyads and triads as the primal form

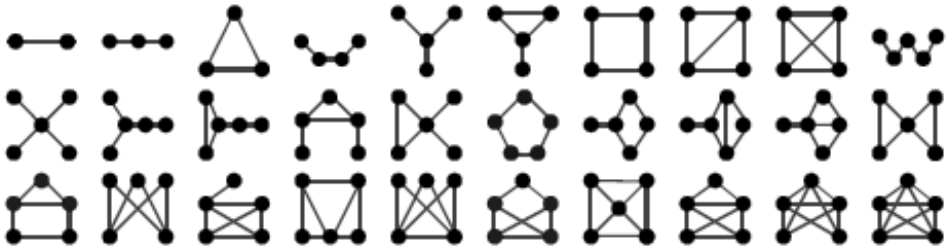


Figure 2.6 – All possible graphlets of size 2 to 5 for undirected graphs

of sociability [132]. More recently, graphlet analysis extended this concept to every pattern of N-entities [?].

Graphlets are defined as small connected *induced*, *non-isomorphic* subgraphs composing any network [91]. In an *induced* subgraph, two vertices linked in the original graph remain linked in the subgraph. For instance, if the original graph is a triangle  we can only induce the simple edge or triangle  subgraph (graphlet). The path of length 2 has all vertices of the original graph but misses an edge and is, therefore, not a possible graphlet.

Figure 2.6 shows all graphlets of size 2 to 5, for undirected networks. Graphlets counting shows that graphlets are not found in a uniform distribution in social networks [?], thus revealing that social networks do not have the same structure that random networks. Precisely, entities in real-world networks tend to agglomerate into groups (also called clusters) where entities in the same groups interact more between them than with entities from other groups [?]. From a sociology perspective, it means that people tend to interact and socialize in groups and interact more rarely with other people from outside groups. These groups are often referred to as *communities*, and many algorithms have been proposed to find these automatically [?].

However, networks concepts, measures, and algorithms have not been used only to study groups, organizations, and societies, but also to focus on separate specific individuals. Indeed, two distinct methodologies emerged through the history of SNA: the structuralists and the school of Manchester [39, 43, 85].

The structuralists are interested in observing the relational structures and patterns forming a network, to make parallels between them and the social behaviors of actors in real life [79]. They think the positions of the persons in the network and the relational patterns they are part of reflect well the social activities and behavior in real life. Accordingly, sociologists in this school usually study organizations and specific groups—such as institutions, companies, families—and want to explain their functioning through the description of the internal shapes and structures of the networks. Thus, they try to construct networks that exhaustively model all the interactions between the actors constituting the groups, as missing links would misrepresent the reality of interactions.

In contrast, the school of Manchester constituted by anthropologists focuses on studying specific individuals and all their interactions in the different facets of their lives and through time. They typically want to explain certain behaviors and social characteristics of individuals by their relationships and interactions in all their complexity and highlight the influence of different social aspects between them in one's life. One famous example is Mayer's study on austral Africa rural migrants going to cities [86] where he showed that the integration of urban mores and customs were directly correlated to the persons' relationships networks in the city. Xhosa[3] people still interacting with rural people of their village in the city were less changing their customs. This school of thought typically relies on the concept of ego and multiplex networks [39]. Ego networks are networks modeling all the direct relations of one central node—in this case, a person—including the relations existing between the persons of this small network. They typically try to model the different types of relationships of a person, like their family, work, and friendship ties, and study them through time. By studying the ego network structure of someone, sociologists of this school try to leverage explanations on other social aspects of the persons like their social status, job, and gender. It is also common to compare several ego networks to make correlations between the social relationships of individuals and other interesting social categories [**?**].

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

2.3.3 Historical Social Network Analysis

History started to use concepts and methods from SNA in the 1980s [149] in response to quantitative history, and to develop historical approaches—like *Microstoria* [49]—that focus on the study of individuals and small groups through the lens of their interactions and relationships directly extracted from historical documents. Beforehand, historians were already describing and studying relational structures such as families and organizations with qualitative methods and with classical taxonomies, without necessarily studying the relational aspect of these concepts. Network research allowed them to model those relational entities more thoroughly using networks, thus allowing them to make new observations that it was not possible to make without taking into account the relational structure of these entities [23]. Since then, HSNA—a term coined by C. Wetherell in 1998 [149]—has been applied by historians to study multiple types of relationships, like kinship [58], political mobilization [84], administrative and economic patronage [97]. If these approaches fall under some of the same critics as quantitative history [80] like lead-

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

ing of trivial conclusions, it still led to classical work and interesting discoveries, such as the study of the rise of the Medici family in Florence in the 15th century by Padgett & Ansell [105], or Alexander & Danowski study on Cicero's personal communications [**?**]. In this work, they modeled the communication of Cicero into a network using 280 letters written by him between 68 B.C. and 42 B.C. It allowed them to study the relationships between knights and senators—which is a subject of interest in Roman history—and concluded that knight-knight interactions were very rare compared to senator-senator and senator-knight interactions. Cicero communication network is illustrated in Figure 2.7.

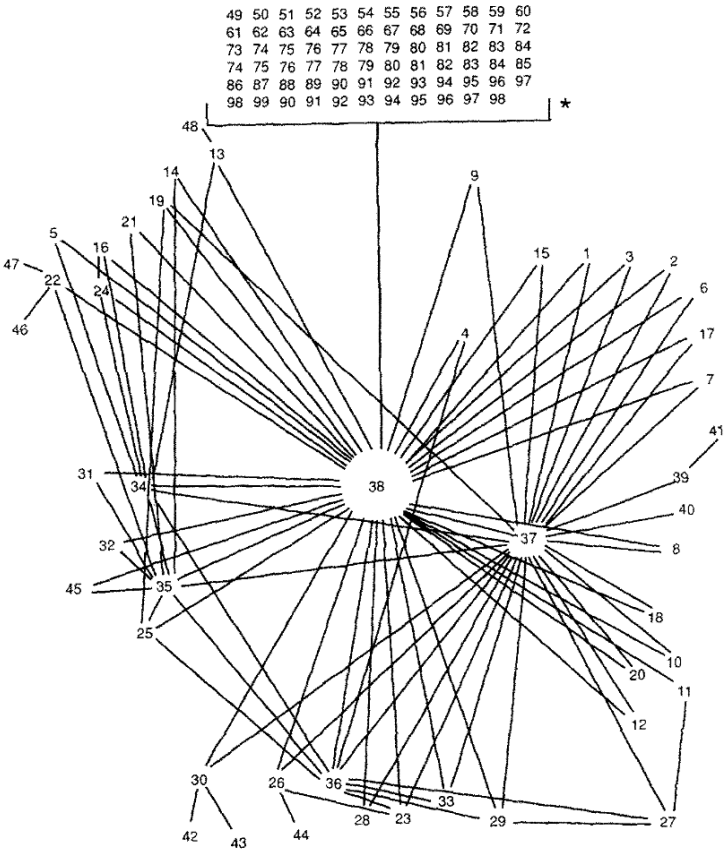


Figure 2.7 – Cicero personal communication network represented with a node-link diagram. Image from [**?**]

Several historians are using and continuously reflecting on HSNA methods [80] which can be very effective to study relational historical phenomena [72]. Moreover, historians rarely rely on a single approach when studying an era or phenomenon, they mix methods and tools from several domains of social and natural sciences with their own practices [105, 109].

2.3.4 Network Modeling

Constructing a network from historical documents, which can vary tremendously in their formats and structures, is not a trivial task [2].

The most straightforward and well-known approach consists in constructing a network based on a simple graph such as in Equation 2.1, where the nodes are the persons extracted from the documents and a link is created when a mention of a social relationship is mentioned is the document (or if they appear in the same document) [?, 80]. This allows to have a simple network to visualize and analyze, but it does not always reflect the sociological complexity of information contained in the documents. HSNA network models have evolved over time to better take into account concrete properties of social networks, such as the importance of actors or relations with weighted networks, multiple relationships with multiplex networks, dynamics of relations with dynamic networks.

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

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

Historians, demographers, sociologists, and anthropologists have also been designing specific data models for their social networks, based on genealogy or more generally kinship [59]. For genealogy, the standard GEDCOM [47] format models a genealogical graph as a bipartite graph with two types of vertices: individuals and families. This format also integrates an "event" object but it is diversely adapted

in genealogical tools. The Puck software has extended its original genealogical graph with the concept of "relational nodes" to adapt the data model to more family structures and to integrate other social relationships for anthropology and historical studies [58].

When creating a network, sociologists and anthropologists can use direct observations of the real world, which is not the case for historians who only have access to biased and partial sources. Indeed, the documents historians inspect are often produced by the political and economical elite of the time, and include the subjective view of the authors, especially for literary sources (letters, journals, books, etc.). Historians therefore need to take a critical view on the sources by acknowledging the position of the authors of the documents compared to the rest of the society, and include it in the analysis [81]. Furthermore, the partiality of the sources often do not allow to have access to all possible relationships types of individuals. For example, if many formal relations can be extracted from official documents such as marriage acts and census, informal relations such as friendships can exist without leaving any written trace [80]. Even for official relationships such as parents and witnesses, there are high chances for missing documents, which do not allow to make too general and finite claims, such as "X is always the case" or "XX is never the case" [**?**]. Social historians therefore have to take into account the partiality and ambiguity of their sources into their analysis, in order to avoid including the bias inherent to their data into their high level historical conclusions.

2.4 Social Network Visualization

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

2.4.1 Graph Drawing

Sociologists rapidly saw the potential of graphically showing relationships between individuals, to better comprehend the underlying social structure and communicate their findings [42]. Moreno elaborated sociograms to visually show friendships among schoolchildren with circles and lines to respectively show children and friendships ties [94]. This type of representation—commonly called node-link diagram—is the most widely used in social sciences, as it is rapidly understandable and effective for small to medium-sized networks which are predominant in social sciences. 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

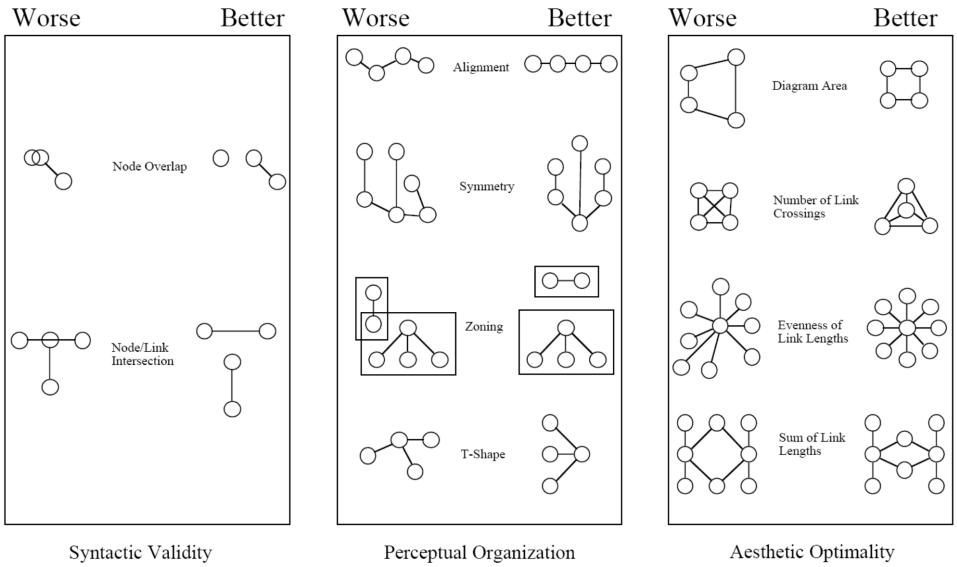


Figure 2.8 – Different criteria are proposed to enhance node-link diagram readability. Image from [75]

the number of edge crossings, the variance of edge length, orthogonality of edges, etc [24, 75]. Figure 2.8 shows some of these metrics, synthesized by Kosara and al. [75]. In Figure 2.5 we can see the difference in readability between the original manual layout (left) and an automatic one (right). Automatic layouts which aim at optimizing readability metrics give clearer diagrams. The number of edge crossings is often considered the most important measure, but finding a drawing with the optimal number of crossings is an NP-Hard problem, meaning that heuristics are needed for most real-world use cases. A large number of algorithms have been designed such as force-directed ones [?], modeling the nodes as particles that repulse each other and are attracted together when connected with a link that can be seen as strings. Other visual techniques have been proposed to represent networks such as matrices, circular layouts, and arcs, but are less used in social sciences [88]. Still, Matrices have been shown to be more effective than node-link diagrams for several tasks such as finding cluster-related patterns, especially for medium to large networks [?, 48].

As social scientists are using more complex network models such as bipartite or temporal networks, more sophisticated representations are needed. The visualization community developed new representations to visualize other network types such as dynamic hypergraphs with PAOHVis [142], clustered graphs with NodeTrix [64] (illustrated in Figure 2.9), geolocated social networks with the Vistorian [128], and multivariate networks with Juniper [102]. However, these new network representations take time to be adopted by social scientists who rarely use them.

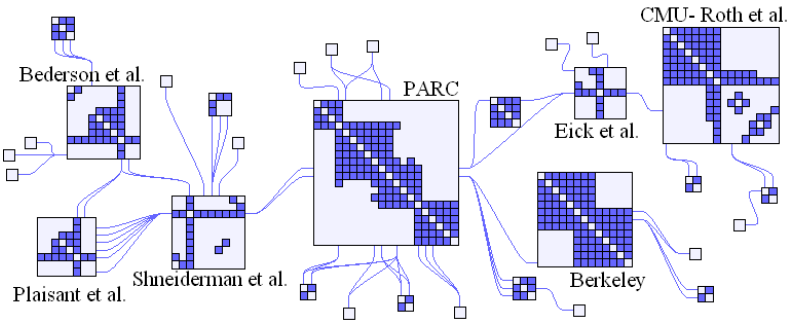


Figure 2.9 – NodeTrix system showing a scientific collaboration social network with clusters. Each cluster is represented as a matrix, Image from [64].

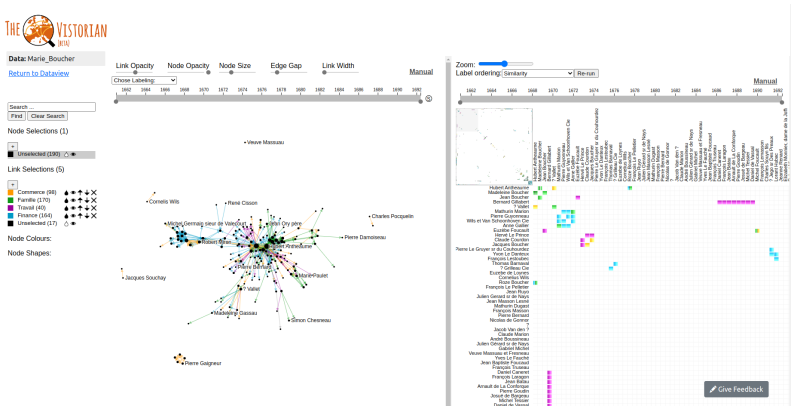


Figure 2.10 – Vistorian interface [128] used to explore a historical social network of business trades in the 17th century, with a coordinated node-link diagram and a matrice view.

2.4.2 Social Network Visual Analytics

Social scientists use visualization and analytical tools to gain insight on the structure of their finalized network data. Most widely used tools are Gephi [5], Pajek [98], Ucinet [**?**], and NodeXl [133] which provide node-link diagrams, implementations of network measures, algorithms, and clustering capabilities. Other SNA visualization tools have been proposed in the past such as Visone [**?**]. However, those softwares often do not include interaction and direct manipulation mechanisms, making the analysis more tedious for social historians and pose usability issues. In contrast, the Vistorian [128] let social historians visualize their network with multiple coordinated views (node-link, matrice, arc-diagram, and map), filters and direct manipulation, but do not integrate analytical options. Figure 2.10 shows the Vistorian interface used to explore an historical social network.

I propose a classification of all those softwares in 2.1, which illustrate that most used software are analytical oriented (Pajek, Ucinet, Gephi, NodeXl) while

	Visualizations	SNA Measures and Models	Clustering	Filtering	Interaction/Direct Manipulation
Pajek	■□□	■■■	■■□	■□□	□□□
Ucinet	■□□	■■□	■□□	□□□	□□□
Gephi	■■□	■□□	■□□	■□□	■□□
NodeXl	■□□	■□□	■■□	■□□	□□□
Vistorian	■■■	□□□	□□□	■■■	■■■

Table 2.1 – Comparison table of most widely used visualization and analytical tool for HSNA. Visualizations: number of different visualization techniques, layout, and interactions. SNA and Models: Number of proposed SNA measures and algorithms. Clustering: Number of proposed clustering algorithms. Filtering: Possibilities of filtering according various criteria. Interaction/Direct Manipulation: Number of possible interactions mechanisms directly applicable on the visualizations.

the Vistorian is an interactive visualization tool. None of those software therefore fully correspond to the Visual Analytics label [**?**].

If analytical methods such as the computation of network measures, triad computation, or clustering provide a good framework to describe the structure of a network and link it to sociological explanation [127, 148], many social scientists such as historians are not trained in computer science and mathematical methods and therefore have trouble to first use those methods without guidance, but also interpret their results. This is particularly the case for black-box algorithms such as for clustering tasks: they usually end up trying several algorithms until they stumble upon a satisfactory enough solution [**?**].

Moreover, preparing and importing the data into visual and analytical software is complicated, as the annotation and network modeling process have not been globally formalized and every historian use different methods, formats and models. Many users do not succeed in importing their data in those systems without concrete help and guidance [2, 128] due to mismatches with data models, formats, or data inconsistencies (null values, white spaces, etc.). If they succeed visualizing their data, it often shows them these inconsistencies or errors such as duplications of entities or wrong attribute values. In other cases, they realize the network do not allow them to answer their sociological questions [80]. It leads to continuous back and forth between their analysis process inside the analysis tool they are using, and their annotation/modeling process, to correct errors or modify annotations. Interestingly, the network model choice plays a crucial role, as a simple network model representing only the persons (as it is often the case) makes it harder to trace back to the original documents containing the annotations from the network entities. Yet the majority of SNA systems enforce simple network models, making this retroactive process harder.

Some interfaces not primarily designed for social scientists incorporate data models encapsulating document representations, such as Jigsaw [136] which is a

VA systems using textual documents as a data model, originally developed for intelligence analysis. It allows an analysis of the documents and their mentions of entities (persons, locations, institutions, etc.) through multiple coordinated views. Using such model allow to rapidly see errors and inconsistencies in the documents annotations that the user can directly correct, while still following complex analyzes.

Finally, more work is still to be done on social network VA tools, to provide more guidance and power to social scientists while doing their analysis, and to help them to do easier back and forth between the annotation, network modeling, correcting, and analysis steps, as errors and inconsistencies can cause high variations in the network structure and hence the analysis results [31].

3 HSNA Process and Network Modeling

I describe in this chapter a formalization of the HSNA workflow followed by social historians, to shed light on their process and summarize recurring pitfalls to identify how VA could support them in this workflow. Most HSNA practitioners report on their findings concerning the network they constructed from their sources, but few highlight the process which led to these conclusions from the raw historical documents, even though they have to make several annotation, encoding, and modeling decisions that deeply influence the final analysis [2]. Specifically, social historians can model documents and their content through various network models which have been proposed in the literature. I discuss in depth this step as it impacts the annotation and analysis possibilities, and I give an answer to our first research question **Q1** by proposing to model this type of data with bipartite multivariate dynamic networks. This model satisfies *simplicity*, *document reality*, and *traceability* properties, which we define as critical for social history work from our joint collaborations with social historians and current critics of HSNA [**?**,80,82].

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

3.1 Context

Tools for social network visualization tend to ignore the context in which the networks are produced, where they come from, and the workflow that led from their origin (e.g., documents, polls, interviews, web scraping) to their network form. Yet, practitioners of social history need to inspect and encode their sources in depth using ad hoc methods to generate a network, and sometimes end with errors or simple networks which do not fit their analysis goals [81]. In this chapter, after describing and characterizing the workflow of HSNA [149] from our collaborations with social historians, I explain why and how effective tools for supporting this process should model social networks in multiple steps to support three essential principles: *traceability*, *document reality*, and *simplicity*. These principles emerged from joint experiences as historians and computer scientists while collaborating on multiple projects, and aim at simplifying the HSNA process while enhancing exploration and analysis options and replicability.

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 that they inspect in depth to next extract qualitative and quantitative information allowing them to answer their research questions.

To study relational social structures where individuals influence each other such as families, companies, and institutions, historians rely on HSNA by modeling the social relationships between a set of entities—usually individuals—into a network. However, the process leading to the final network from the raw documents is often linear, and it is common that, when visualizing their network, historians spot errors and inconsistencies in the network structure that they could have fixed if the process was more iterative [2]. Moreover, historical documents are often complex, meaning that the annotation and modeling process can be done in many different ways, concerning what to annotate from the documents [82] and how to model the annotation in a network [23]. 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 modifying/correcting annotations complicated. Indeed, errors and mismatches often occur in the annotation process, for example, due to entity disambiguation problems [31]. 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, I answer **Q1** (how to model historical documents into analyzable networks with the right balance between expressivity and simplicity) by proposing to model historical documents as bipartite multivariate dynamic networks, where both persons and documents are modeled as nodes with attributes and the links represent both individuals' mentions in the documents and their social roles in the event witnessed by the documents (such as witness in a marriage act). 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, network modeling, and social network visualization in chapter 2, we only discuss in this section the related work concerning historians' methodology and workflows.

3.2.1 History Methodology

The essence of the historical discipline is based on a critical approach to 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 an exhaustive set of historical documents [?]. With the development of social and quantitative history, historians now have a panoply of methods to exhaustively extract quantitative data from their sources and analyze it to ground their results in verifiable claims. Many historians criticized this computational aspect of history [?, 83], pointing out that it would lead to errors and missing the core content of historical sources. However, using quantitative approaches and formalisms is not exclusive to having a deep understanding of the documents and their context, nor building a narrative on top of their quantitative analysis. Good historical work can in fact be described as a combination of the two [70], as Tilly says "Formalisms play their parts in the space between the initial collection of archival material and the final production of narratives. In my own historical research, formalisms figure prominently from early in the ordering of evidence to late in its analysis; [...] As it happens, many other historians rush from sources to reasoned narratives without pausing to employ formalisms, or even to reflect very self-consciously on the logical structure of their arguments, hence on what the evidence should show if their arguments are correct" [?]. Historians have a panoply of methods and formalisms they can leverage to ground their narratives in concrete comparable results, such as serial analysis, tabular analysis, classical statistical treatments, and network analysis.

However, formalisms have to be used wisely and with a critical vision of the documents and their context, so as to not fall into simplifications, anachronisms, and errors which are pertinent critics of quantitative history [81,82]. Most historical work leverage several methods in the same study to support their claims through different qualitative and quantitative results [109]. The level of the plausibility of a claim increase or decrease depending on if the different evidence point to similar results or not. Similarly, historians often work on small populations or specific individuals—as it is the case with microhistory studies [49]—which can result in complications for generalization. Only after studying several similar individuals or groups, historians are able to generalize and point to exceptions. For example, by comparing several Jewish commercial communities in Europe during the first half of the 18th century, Trivellato has been able to generalize what is common to those groups (they have been trading between them and with outer ethnic groups) and what is specific to each (such as their business strategies) [?].

3.2.2 Historian Workflows

Many quantitative methods and formalisms are available for historians to inspect their sources in the aim of making historical claims. Several textbooks describe and explain to social scientists and students who do not have formal computer science training what consist these methods (statistical regression, Chi-

squared test, network analysis, etc.) and how to practically use those with software and programming language [?]. However, the process leading from the sources to the numeric artifacts (a table, a network, a timeline) has not been described thoroughly in the literature, especially with concrete examples, and is often not presented in scientific publications of concrete use cases. Yet, the process leading from the documents to analyzable data requires social historians to make several annotations, encoding, and modeling decisions, concerning *what* to extract from the source and *how* to encode it. This process is tedious and requires data acquisition, annotation, encoding, and modification with continuous back and forth between the different steps [2]. This is a critical step as it can lead to simplifications, anachronism, distortion, or data that do not allow to answer original or new hypotheses [70,81]. Lemercier and al. give guidelines on how to encode information from historical documents to prevent introducing bias, by having a critical view of the documents [82]. They emphasize the importance of the input phase of research and advise copying the first documents by hand while characterizing them in the most exhaustive and factual way, without imposing categorization. This explorative step let historians familiarize themselves with the content of the document, leading to a better view of what to encode to answer their research questions and sometimes to the formulation of new hypotheses. For example, in their project on the social and geographical trajectories of Jews in Lubartów [?], a village in Poland, the team noted the mean of writing inside the register documents (pen, pencil, ink, etc.) they were inspecting. This information allowed them to conclude that the inscription "expelled" written in pencil was probably added during World War II by Germans to denote exported Jews in the extermination camps. When applying network analysis, historians often create simplistic networks which allow them to answer specific research questions, but often loose this type of information related to the documents. Cristofoli discusses the network modeling problem when following a network analysis and highlights the fact that the same historical documents can be modeled in different ways [23], which can result in mismatches between the network shape and the research questions. Dufournaud presents her quantitative and network workflow when studying the economical role of women during the 16th and 17th centuries in the city of Nantes, which she splits into three steps: data collection, data processing, and data analysis [34].

3.3 Historical Social Network Analysis Workflow

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

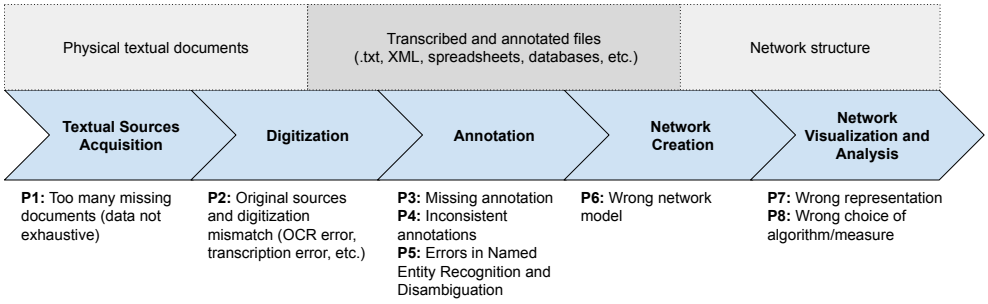


Figure 3.1 – HSNA workflow is split into five steps: textual sources acquisition, digitization, annotation, network creation, and network visualization/analysis. Practitioners typically have to do back and forth during the process. I list potential pitfalls for each step.

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

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

The Text Encoding Initiative (TEI) [22] is an XML vocabulary and a set of guidelines typically used to encode and annotate documents, and the events happening in these documents (unclear parts, gaps, mistakes, etc.). It is also used for historical texts and to generate social networks [35, 128]. Unfortunately, the guidelines are not meant to define a canonical annotation and different persons can interpret the guidelines in different ways, leading again to inconsistent annotations of corpora (P4) and to errors or distortions in social networks derived from these annotations.

3.3.4 Network Creation

Historians construct one or multiple networks from the annotations of the documents. Typically, all persons mentioned are annotated and are 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)**, such as the creation of network structural artifacts when using network projections [23]. More complex models have been proposed in the literature such as weighted, dynamic, bipartite, and layered networks, but can be hard to manipulate and visualize.

3.3.5 Network Analysis and Visualization

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

3.4 Network modeling and analysis

Historians typically construct one or several networks from their annotated documents that they will visualize and analyze to validate or find new hypotheses. As the processing steps of the workflow are often not transparent (digitization, annotation, network modeling), it can be difficult for the reader of an HSNA study to understand how the network has been constructed, what it represents, and to trace back the network entities to the original sources [34]. Moreover, visualizing the network very often highlights errors and artifacts of the annotations, along with potential mismatches between the network model and the analysis goals. Historians then have to correct or change their annotations, even though it is a very tedious and demanding process to repeatedly switch back and forth between the network and the annotated documents. Several network models make the task harder as they do not directly represent the documents, and it is thus difficult to relate a network entity to a specific document and annotation. Therefore, we believe that more visual analytics tools should support social scientists in annotating and modeling their documents to make the HSNA process less linear by allowing easier back and forth between the annotation, modeling, and visualization steps. Network models satisfying *traceability*, *document reality*[1] 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 data modification goals.

3.4.1 Network Models

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

- **Simple Networks [149]:** According to their research hypotheses, historians select and merge document information to build a specific relationship between individuals. They analyze this simple network structure with SNA tools and produce network indicators and node-link visualizations. It is often difficult to connect the results to the original sources.
- **Co-occurrence networks [123]:** Only the persons are represented as nodes, and two persons are connected with a link when they are mentioned in the same document (or section). This is a simple model and one of the first to have

[1] We chose the term "document reality" over simply "reality" after a conversation with a historian to highlight the fact the historical documents do not describe factually the reality and reflect the subjective bias of the context in which the person wrote them [70]. The content of the documents, therefore, has to be modeled by taking into account this context, which can reveal interesting behaviors and structural patterns. See [82] for specific examples.

been used in SNA and HSNA. The major drawback of this model is that it does not take into account the diversity of social relationships, as every link is identical. It can work well when only one type of social relationship is studied like a friendship network [95]. However, historical documents rarely mention only one type of relationship and this model is thereby very limiting for HSNA.

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

to be visualized, with no generic system to support historians in taming their high complexity.

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 *document 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, resulting in simplifications and distortions of the structural patterns. For example, modeling only co-occurrences of persons in documents remove the variety of social relationships these mentions can refer to [80]. 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, meaning that the data is harder to correct and that analyses are less replicable. On the other side, multilayer networks and knowledge graphs 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 *document reality* properties can be hard to achieve. The lack of simplicity, therefore, results in difficulties for visualization and analysis, especially for social scientists.

3.4.2 Bipartite Multivariate Dynamic Social Network

Historical documents are well modeled by bipartite multivariate dynamic networks with roles, that can be formalized as

$$G = (V, E, B, R, T, L) \tag{3.1}$$

where V is the set of vertices, E the set of edges, and $B = (person, document)$ the set of node types. Each node $u \in U$ is defined as

$$u = (u_{id}, b, a_u) \tag{3.2}$$

where $b \in B$ is the type of the node and a_u is a tuple of the attributes (or properties) of u such that

$$a_u = (a_i, ..., a_n) \tag{3.3}$$

with $a_i, ..., a_n$ the attributes of the node u defined on their domains $A_i, ..., A_n$. We do not impose constraints for person nodes, but document nodes always have a time and location such that when $b = document$ then

$$a = (t, l, a_i, ..., a_n) \tag{3.4}$$

with $t \in T$ is the time of the event witness by the document and $l \in L$ its location.

Similarly, each edge $e \in E$ is defined as

$$e = (u, v, r, a_e) \tag{3.5}$$

with u, v the vertices connected by e such that $b_u \neq b_v$, $r \in R$ the role of the person mentioned in the document and a_e the attributes tuple of e such that

$$a_e = (a_i, ..., a_n) \tag{3.6}$$

with $a_i, ..., a_n$ the attributes of the edge e defined on their domains $A_i, ..., A_n$.

The model has therefore the following properties:

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

Links and Roles: A link models the mention of a person in a document. **Each link has a type corresponding to the role of the person in the document**. For a marriage act, the roles include *wife*, *husband*, *witness*. This is a key aspect of our model since it clarifies the relationship between the persons within an event. In contrast, Jigsaw [136] does not consider the roles.

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

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

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

One of the main benefits of this model is that the document nodes represent both the physical documents and the events the documents refer to. For example, concerning marriage acts, the document nodes represent both the physical documents with their texts but also the marriage events with their characteristics modeled as attributes (time, location, etc.). Therefore, social historians can use this model to store, process, and annotate 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 *document reality* of the social relationships mentioned in the sources as no projection or transformation is applied.

Visualization tools using this model can focus on the topology of the network, and/or the attributes which I express here in the format of tuples, commonly used by databases and visualization systems [?]. However, it has to be taken into account that if the attributes extracted from the historical documents are related to vertices and edges independently to the topology of the network, it can be appropriate to compute vertices and edges measures—such as the centrality—and store them similarly to the other attributes, especially so that visualization systems can leverage the same interactions for both. In that case, these types of attributes are directly dependant of potential topology changes in the network (in the case of subgraph extraction or network modifications interactions for example).

3.4.3 Examples

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

1. Analysis of the social dynamics from **construction contracts in Italy in the 18th century [25, 101].** The corpus is made of contracts for different types of constructions in the Piedmont area in Italy. People are mentioned under three different roles: *Associates* who are in charge of the construction, *Guarantors* who bring financial guarantees, and *Approvers*, who vouch for the guarantors. Documents contain information about the building site, the type and materials of constructions, and the origin of the people.
2. Analysis of migrations from the **genealogy of a French family between the 17th–20th centuries** [unpublished work]. The corpus is made of family trees referring to several document/event types: birth and death certificates, marriage acts, military records, and census reports. The roles are different for each event type and consist of *children, father, mother* for the birth events, *deceased* for the death event, *spouse* and *witnesses* for the marriages, and *family member*s for the census events.
3. Analysis of migrations from Spain to Argentina through the **marriage acts at Buenos Aires in the 17–19th centuries [96, 122].** The corpus is made of summaries of marriage records that mention the spouses and the witnesses of the wedding. The origin, date of birth, and parents' names are specified for both spouses.
4. Socio-political analysis of **migration of ethnic Germans from communist Romania to West Germany in the 20th century (ongoing work) [32].** The corpus is made of administrative forms that mention persons requesting to migrate, along with the persons they want to join, and the administrative persons of the ministry in charge of the forms. The family members of the

aspiring migrants are also mentioned in the forms, with their respective dates of birth.

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

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

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

Finally, these examples show that when working with multivariate networks, using projections to create unipartite networks brings a duplication of information. Indeed, if a document mentions information like a date that we model as an attribute, we can store it as a document node attribute using a bipartite model. However, when projecting the network this information appears in the links as many times as there are persons mentioned in the document minus one and often more. For example, in the example #1 in Table 3.1 the time is stored in $\sum_{i=1}^{4} i = 10$ links in the co-occurrence model and in 9 links in the multiplex unipartite model while it is only stored once as a document node attribute in the bipartite model.

3.5 Applications

Several tools have been designed for visualizing dynamic bipartite networks that can also be considered dynamic hypergraphs [108, 142], but few incorporate attributes. Moreover, the vast majority of visual analytics tools are solely focused on the analytical part of the data, meaning that the link between the original documents and the hypergraph abstraction is often broken. Social scientists therefore

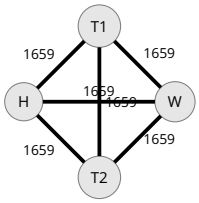
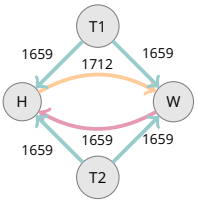
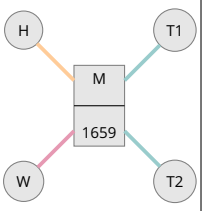
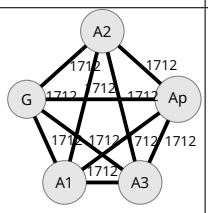
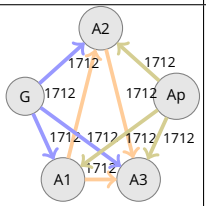
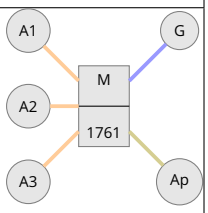
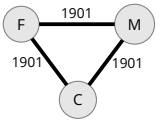
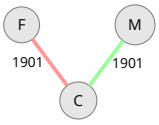
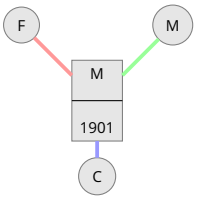
Original Document	Co-occurrence	Unipartite representation	Bipartite
20-4-1659 : Capitán Alonso MUÑOZ de GADEA , con Da. Francisca CABRAL LEAL de AYALA . Ts.: Agustín Gayoso , y Juan Guerrero. Al margen: "fue Oficial Real" , (f. 9v). Husband Wife Witness			
1712: Construction of a church in Torino. Associates: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

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

always have to do many back and forth between the visual analytics tools and their original documents and the annotation/modeling processes. More visual analytical tools should thus incorporate the textual documents in their data model similarly to Jigsaw [136], as it would allow tracing the entities of the network back to the original documents more easily. Mechanisms to 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 correct errors and inconsistencies in the annotations and propagate them in the visual analysis workflow. For example, the Vistorian [128] now lets users modify and correct 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 *document 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 3.2for the two cases where marriage acts mention birth information and the case where only marriage-related information is present in the document.

3.7 Conclusion

HSNA is a complex process that starts by collecting historical documents and ends with elaborating high-level sociological conclusions. Historians support their conclusions by modeling individuals' social relationships extracted from the doc-

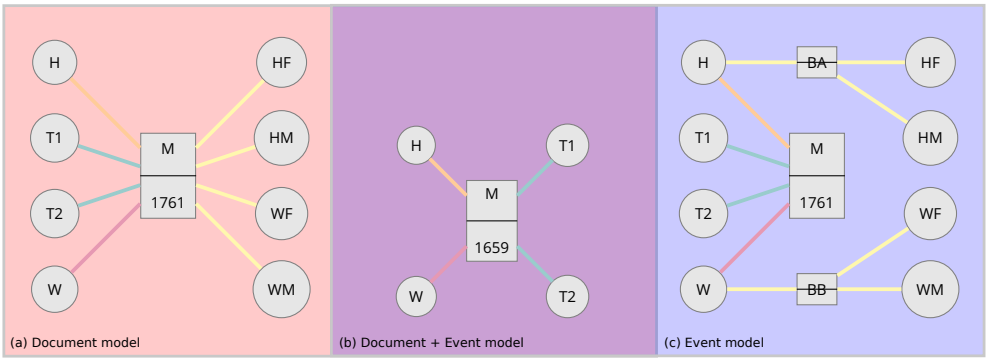


Figure 3.2 – bipartite multivariate dynamic network modeling for two cases of marriage acts of example #3. Some marriage acts mention the parents of the spouses, which is a relationship different than the marriage in itself. This case can be modeled using a document model (a) or an event model (c) by 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.

uments and analyzing them through network visualization and analysis methods. Most historical work do not provide details on how they constructed their final network, even though it is a complicated and tedious process that can result in many biases and distortions if not done carefully [2]. We shed light on this process by dividing it into 5 steps and describing recurrent pitfalls we encountered in our projects and collaborations. More importantly, I explain why this process should be done following the principles of *document reality*, *traceability*, and *simplicity*, to avoid biasing the analysis, allowing to go back to the original source at any point of the workflow for easier corrections and replicability, 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. I 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, thus answering **Q1**. 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, I showed that bipartite multivariate dynamic networks allow to model well historical documents, with *simplicity*, *reality* and *traceability* properties. However, no visual tools currently exist to specifically explore and manipulate this type of data. In this chapter, I propose a VA interface aimed at exploring historical documents modeled as bipartite multivariate dynamic networks, for historians to be able to reflect on their data and encoding, and to follow in depth-analysis. I answer **Q2** by analyzing tasks and questions historians have on their data and providing interactions mechanisms which would allow them to answer their historical questions while finding errors in their data.

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

4.1 Context

Social scientists such as historians aim at characterising the structure and dynamics of social groups of interest, on a region and period of time 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 network analysis approach [149]. For this, they need to encode their documents to extract the persons and any other useful information in the text and transfer it into a structured file or a database. Social scientists can then explore, validate, or refute their hypotheses by visualizing and analyzing the network structure and the connectivity patterns between the entities of the resulting network.

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

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

In addition, historical documents include time and geospatial information corresponding to the date and location of the events they refer to, and potential additional information on the mentioned individuals such as their gender, profession, and date of birth. These are often essential to understanding underlying social phenomena, as time, space, and classical social categories play an important role in sociological structures and dynamics [80]. For these reasons, as we discussed in chapter 3, historical sources and the underlying social events 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 represent historical documents with simplicity, traceability, and document reality, i.e., the relationships appear as they are mentioned in the documents without distortions implied by projections [23].

More complicated models exist such as knowledge graphs [**?**], which are very expressive but hard to manipulate, especially for social scientists. In contrast, most visual and analytics tools widely used by social scientists such as Gephi [5] and NodeXL [133] enforce too simplistic network models and only provide limited interactions for exploring the network data, even if they provide the computation of several network measure. This results in many social historians ending their analysis by plotting the network using a node-link diagram—which is hard to read with dense networks—, and identifying the most important actors with the help of centrality measures [81]. Lemercier & al describe this phenomena the following: "Network graphs of the "spaghetti monster" variety are a case in point. Often, in historical papers, they are used to show that a network is dense (and that the author has mastered the new technology). The narrative then comments on the individuals identified as central in the network. This approach can indeed be quite interesting, but it is hardly the only possible use for network analysis. [81]". VA tools guiding social scientists towards more complex exploration with the help of interaction mechanisms are therefore needed.

In this chapter, I present a VA system to explore and analyze Bipartite Multivariate Dynamic Social Networks, with the aim of supporting more complex historical analysis based on easy-to-use interactions, but also potential data correction. I

elaborated the tool based on four collaborations with social scientist colleagues. I 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—corresponding to specific groups or individual exhibiting intriguing behaviours—or in comparing several subsets of the network, in terms of network measures, attribute distributions, and their overlaps.

I hence 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. I 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. By easy-to-create queries, social historians are able to 1) detect erroneous patterns and reflect on their encoding process and 2) find relevant patterns which can answer their historical questions. For this, I rely on the Neo4j graph database [100] and its query language *Cypher*. Most visualization systems offer dynamic queries to hide the complexity of query languages. However, using a rich data model, some queries are easier to refine using scripting than dynamic queries. I 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. Furthermore, they can export their query, the associated results, and its history at any point to share it 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. I validate the query and comparison system with a formative usability study and I demonstrate ComBiNet can be used to answer sociological questions by describing in depth several real-world use cases.

After the related work section, I describe our design process in §4.3, present the system ComBiNet in §4.4 with the design of the visual query and comparison features, and present four use cases demonstrating the utility of the system in §4.5 showing it can be used to explore complex historical data and allowing historians to answer several of their questions using queries and comparisons. I finish with the results of the formative usability study in §4.6.

4.2 Related Work

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

4.2.1 Graphlet Analysis

One of the inspirations for this project came after participating in the 2020 VAST challenge[1] where our team used graphlets to measure similarity between several networks [139].

Graphlets are small connected induced, non-isomorphic subgraphs composing any network (see §2.3.2 for mor details). They were first introduced by Milo et al. [91] to explore the structural differences between biological networks, but they are now used in several disciplines involving networks such as sociology [**?**].

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

Instead of counting every graphlet occurrence and interpreting them with sociological meanings, it appeared more efficient to let social scientists finding specific patterns to answer questions they already ask themselves on the data.

4.2.2 Visual Graph Querying

Graph pattern matching is an important task in SNA, which consists in finding a subgraph of interest in a larger graph [**?**].

Several scripting languages, such as R [114] and Python [143], have been extended to support the exploration of social networks using specialized libraries such as igraph [27] and NetworkX [57], and provide functions for graph pattern matching. However, social scientists are often challenged to use scripting languages and programming, as they often do not have formal training in such technologies.

Graph databases allow to store and manipulate network data with the use of query languages, such as the Cypher language for Neo4j [100]. To lower the complexity barrier of their usage, several visual graph query systems have been developed to allow analysts to rapidly build and refine their queries visually. Some systems hide the scripting language such as GRAPHITE [20] and VERTIGo [28] that allow specifying a graph query as a node-link diagram that the user creates interactively. Shadoan and Weaver [129] use a similar concept with hypergraphs

[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.

to filter multidimensional data. Other systems, such as VIGOR [111] only visualize the query after it has been written by the user with a scripting language and do not allow direct interaction of the visual representation of the query. All these visual query systems are limited to topological queries with constraints on the vertex and edge types and do not allow to make constraints on other dimensions such as attributes and time associated with vertices and edges.

4.2.3 Visual Graph Comparison

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

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

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) plus collaborative and reproducibility capabilities [115]. For example, the VisTrails system allows users to reproduce their visual analyses by providing an executable history graph of their actions, [17] while GraphTrail provides provenance tools to ease collaborative analysis [36]. Provenance can also be beneficial for visualization designers and researchers, as it gives them a tool to understand users' behaviors [7, 12] and evaluate/improve visualization systems [117]. All the reasons and concrete implementations of provenance are discussed in depth in Xu's survey [150].

4.3 Task Analysis and Design Process

I designed the ComBiNet tool in collaboration with social historians who wanted to follow a network analysis on their historical semi-structured documents, that are well modeled by bipartite multivariate dynamic networks. I 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 that I used to design and implement the interface, with continuous feedback from our collaborators.

4.3.1 Use Cases

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

1. Analysis of the social dynamics from **construction contracts in Italy in the 18th century (141 documents, 272 persons) [25].** The corpus is made of contracts (manuscript documents) for different types of constructions in the Piedmont area in Italy. People are mentioned in three different roles: *Associates*, who participate in the construction; *Guarantors*, who bring financial guarantees; and *Approvers*, who vouch for the guarantors. Along with the time and location of the construction site, documents have a construction type (military, religious, and civil), work type (big work, small work, reparation, transportation, etc.), and material (wood, stone, metal). People also have an origin attribute (the place they come from), manually extracted from the original documents.

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

   Query 1.1 Request all approbators occurrences

   Query 1.2 Request all associates and guarantors occurrences

   Question 2 Are there people mutually guarantors to each other in different contracts? (G, AT, N)

   Query 2.1 Select pairs of people connected each to the two same document, with a guarantor role and any other role

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

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

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

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

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

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

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

Question 6 What are the differences between Torino and Torino surroundings, concerning the types of constructions and actors involved? (G, AT, C)

Query 6.1 Request all documents located in Torino, with the persons mentioned

Query 6.2 Request all documents located in the Torino area, with the persons mentioned

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

Question 7 What is the trajectory of life for a given specific 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 What is the trajectory of live for a family (L, A, N)

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

Question 9 Where are located the main migrations, and at which time do they occur? (G, A, N)

Query 9.1 Select persons with a geolocated birth and death certificate

Question 10 Are there differences in volume and location between migrations in the 18^{th} and 19^{th} centuries? (G, A, C)

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

Query 10.2 Select persons with a geolocated birth and death certificate from the 19^{th} 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 19^{th} century, was there an overall decrease in the social status and professions of persons in the dataset? (G, A, C)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.3.2 Tasks Analysis

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

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

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

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

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

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. Table 4.2 shows a comparison of their data model compared to ComBiNet.

When started, ComBiNet dynamically collects the node types, roles, sub-types, and attributes when reading the network from the database. The interface is constituted of four main panels, split into different views as shown inFigure 4.1: the query and comparison panel (V6, V7, V8, and V9), the bipartite visualization panel (V1), the map visualization panel (V2), and the query results panel (V3, V4, V5). I present in the following the different views, according their visualization or query functions. Comparison features are incorporated in the same views with different comparison mechanisms.

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.

	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.

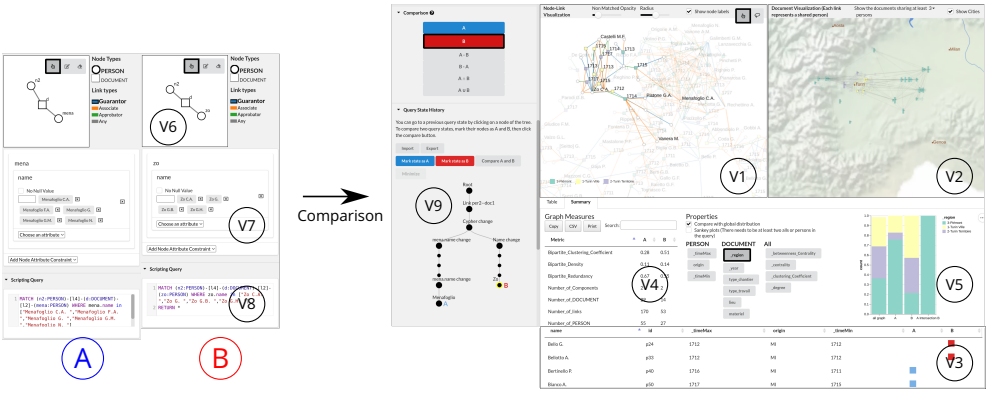


Figure 4.1 – The ComBiNet system used to compare two subgroups of a social network of contracts from [25], extracted with dynamic visual queries. (A) and (B) show the two visual queries created by the user in the query panel using an interactive node-link diagram editor (V6), dynamic query widgets (V7),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

4.4.1 Visualizations

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

V1: Bipartite Node-Link View The bipartite node-link visualization panel shows the network using the DrL force layout from igraph [27] with overlap removal using D3 [13]. Node-link representations are very common in social sciences [5, 24, 98, 133] and were a specific request from our collaborators. In the context of our bipartite model, the persons are represented as circles and the documents/events as squares, while the roles are encoded as link colors. A link models the mention of a person in a document. This view provides an overview of the data by showing the structure of the network (T1.1) and the roles of the persons in their different documents (T1.2). The view also provides pan & zoom and selection interactions for effective navigation. Nodes' labels are displayed (name of person and ids of documents by default) on the canvas with an occlusion-free mechanism which hide nodes with low degree when two or more nodes labels overlap.

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 network, focused on the geospatial information that is very important to social scientists (T1.3).

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

The map has the same interactions mechanisms than the bipartite node-link view. The two views are also 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 (person or document). 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. Selecting a node hence highlights the corresponding row and pushes it at the top of the table. Tables in social network visualization systems have been proven to be efficient and useful for social scientists when exploring their data [9]. It allows them to link the visualization to the network entities more easily, and dive deeper into one entity's attribute values after selecting it in the network. It also makes ranking entities according to various criteria easier and more straightforward.

V4: Graph Measures The Graph Measures view shows measures related to the network and gives insights into its structure to users (T1.1). We report simple measures like the number of persons, documents, links, and components, and more sophisticated bipartite network measures asked by our users, that they can report for their analysis: the bipartite density, bipartite average clustering coefficient, and bipartite average redundancy [**?**]. 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 "All" for both types). 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. Figure 4.2 shows the construction dataset of collaboration #1 where the user selected the *_year* attribute, coloring the documents nodes with

their year in the node-link diagram (left) and the map view (right), revealing for example that most construction occurring in 1714 occurred in Torino and Torino close surrounding. By clicking on the button, the visual encoding and the distribution plot remain selected. This interaction is inspired by the x-ray technique of the Vizster system [61]. Users can follow a first exploration of their data by visually detecting correlations between attribute values and some groups of persons or between attribute values and some specific areas in the map view (T1.2, T1.4, T1.5).

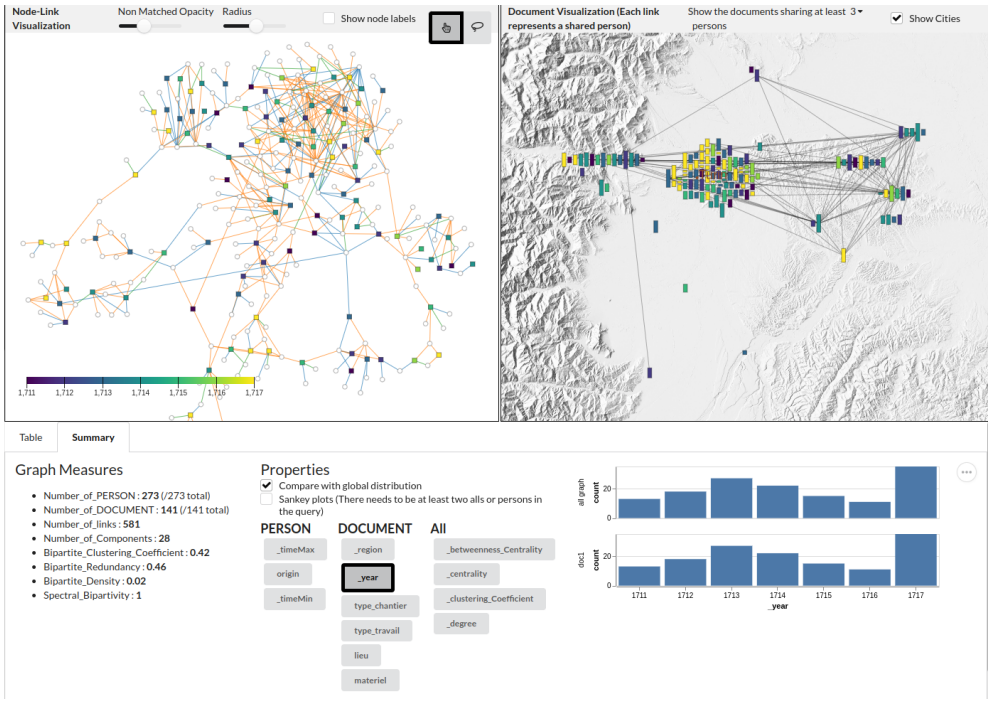


Figure 4.2 – ComBiNet interface with the dataset of collaboration #1. The user selected the *_year* attribute, showing the distribution of document years with a histogram (bottom right), and coloring the documents node on the bipartite view (left) and map view simultaneously (right).

4.4.2 Query Panel

The query panel allows users 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 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 subsection, I describe all the features and interactions allowing ComBiNet to build a query and illustrate them

with questions 2 and 6 of our collaboration #1. Our collaborator wants to *find the persons who are mutually Guarantor to each other in separate contracts (2)* and to know *how Torino and Torino's surroundings differ according to their contracts? (6)*. Figure 4.4 (left) shows the final queries, but first, I explain how to create them.

V6: Node-Link Dynamic Query

The interactive node-link diagram allows the construction of a subgraph query graphically, that represents a topological constraint (T2.1). The query subgraph is built and edited interactively. At each modification, the visual query is converted into a Cypher query and run in the database which return the results. All the matches are displayed in the entities tables (V3) and highlighted in the main visualizations views (V1, V2). 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 or Document).

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 network view: a visual query is a network template with additional attribute consraints. Each role (link type) is rendered using a color (Figure 4.3 left). Users can also create untyped links using the *Any* value, which will match all the existing link types (Figure 4.3 left). Created links can be matched by different selected link types, by checking several possible types for one link. These links are represented by a dashed line with the colors of the possible types (Figure 4.3 middle right). Several links with different types can also be created among two nodes to query a person with more than one role in the same event (Figure 4.3 right). When a node or link is created in the query, it is given an identifier starting with *per* for a person, *doc* for a document, *link* for a link, followed by a number. These identifiers are

used in the attribute constraint panels and the textual query and can be changed through their textual representations.

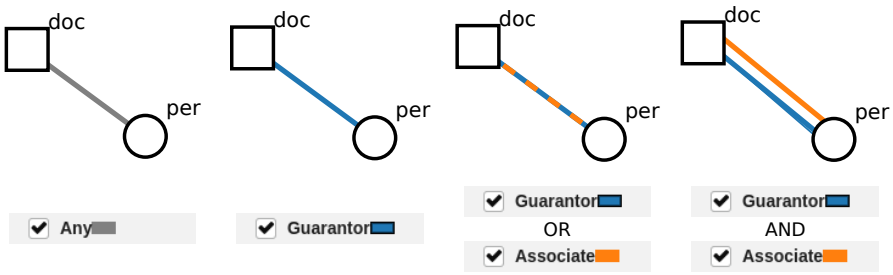


Figure 4.3 – All link creation possibilities: Any link type (left), one selected link type, here guarantor (middle left), the union of several link types (middle right), several links with different types (right)

To find persons who are mutually guarantors in our collaboration #1, we first create one person and two documents using the addition mode and by clicking on the canvas. We then link the person node to the first document with a link that is not typed (Figure 4.3 left), and link it to the second document with a *guarantor* link (Figure 4.3 middle left). We then create a second person node and link it to the two documents with opposite link types. The resulting visual query is presented in Figure 4.4 (a). To answer the question 6 (comparing Torino and Torino surroundings), we start to request all the links in the graph, no matter the type, as shown in Figure 4.4 (b). The database then returns all the links in the graph with their attached nodes. Putting attribute constraints on the location of the contracts will then let us answer the question.

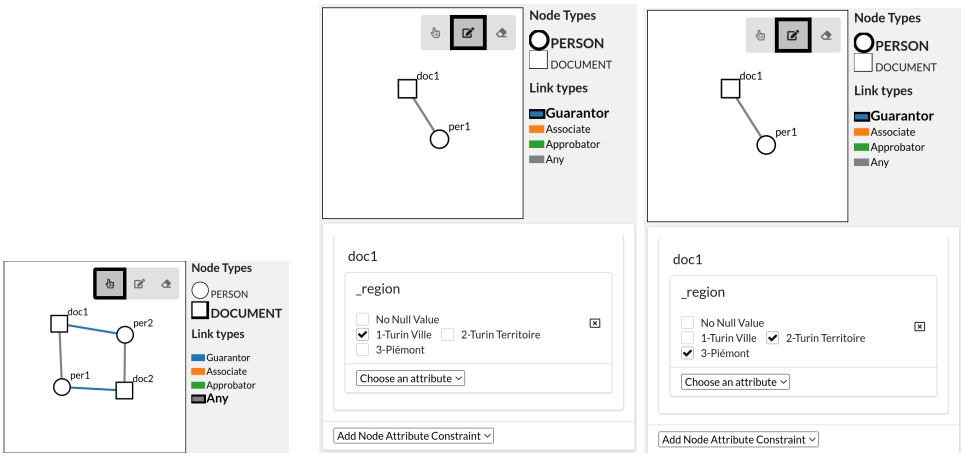
V7: Attribute Constraint Widgets Users can also add attribute constraints (T2.2) on the created nodes with the help of interactive widgets. An input button is created for each node and link identifier from the node-link query panel. It allows to create a dynamic query widget for any of its attributes.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

1. **Numeric constraints** are modeled as range sliders, allowing the selection of lower and upper bounds 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 corresponds to the union of the filters. The three widget types are shown in Figure 4.5.



(a) (b)

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

To answer our collaborator's second question (*how do Torino and Torino's surroundings differ according to their contracts?*), we first filter the documents which are located in Torino (*Turin*). For this, we select the whole dataset by linking a person and document node with an untyped link. Then, we select the id *doc1* of the document of our visual node-link query, and the *region* attribute. It initializes 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 parts of the Cypher query which are not visually expressible appear in Cypher widgets next to the other widgets. The editor supports autocompletion to e.g., help to discover and spell the attribute names. The visual and textual representations are synchronized, meaning that modifying one update the other and return the results in the visualizations, tables, and attribute distributions.

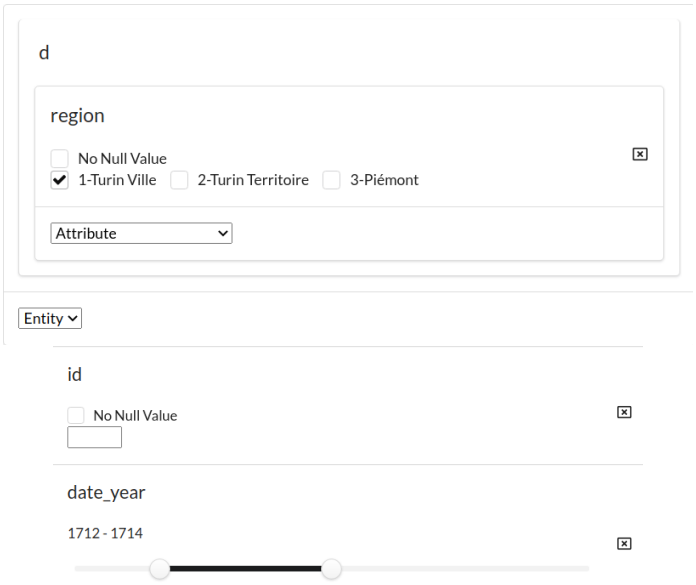
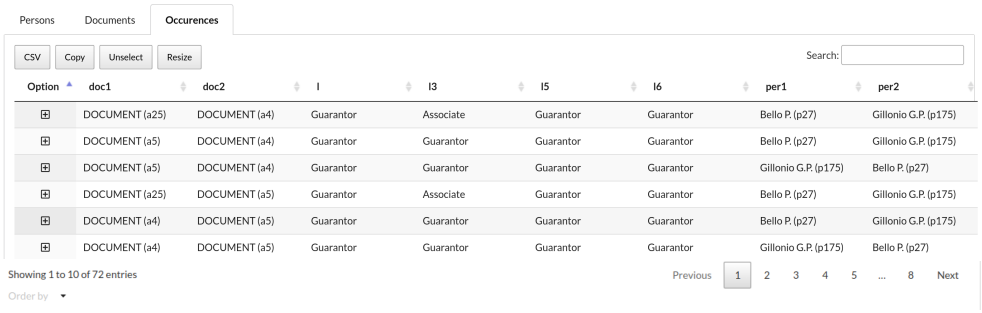


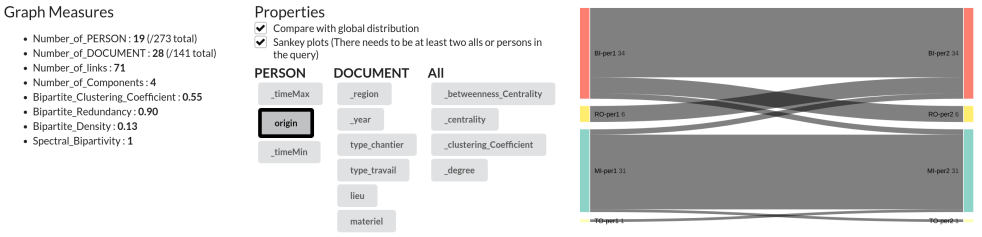
Figure 4.5 – Widget designs for the different attribute types: checkboxes for categorical attributes (top), text input for nominal attributes (middle), and a double slider for numerical attributes (bottom). The categorical attribute example shows the inputs letting users create new constraints for other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

Query Results Each modification of the query, whether from the node-link dynamic query, the widgets, or the Cypher text box, update the two visualization panels (V1, V2), the entities tables (V3), the network 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 occurrence has been found 72 times, meaning that the pattern exists 36 times in the network by taking into account the symmetry of the subgraph query. Users can switch between the three tables in the table view using the tabs. The network measures are computed on the new subgraph formed by the union of all patterns found and updated on the network measures view (V5). Figure 4.6 (b) (left) shows to the user the different graph measures of the subgraph induced by the patterns found. Since some measures can be long to compute, the values are computed iteratively in the backend and shown progressively [41] to avoid blocking the interface. The distribution plots in the attributes view (V6) are updated, showing the values of the entities of the latest constructed query, next to the global distributions.

Attributes Visualization When users select an attribute in the attributes view (V5), its distribution is visualized for the queried entities and the whole network



(a)

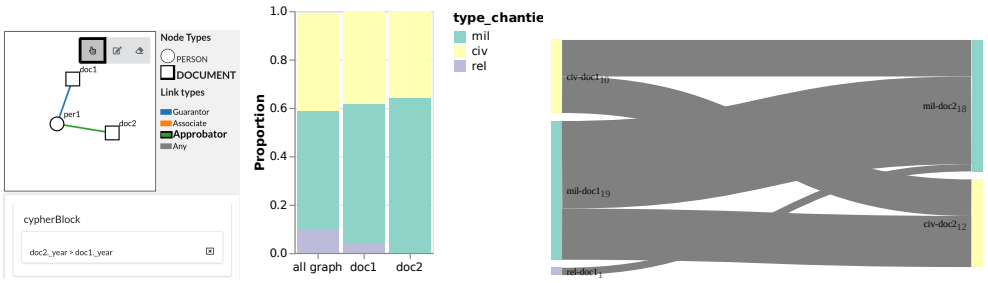


(b)

Figure 4.6 – Results of question 2 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.

with an histogram. 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 nodes have a relationship in regard to time, such as birth certificates, marriage, or 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 views for the results of question 2 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



(a) (b) (c)

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

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 with the Trrack library [29]. Each node of the tree represents a query change, with a description label such as "New Link". It allows to rapidly visualize the succession of filters applied with their refinements. At any moment, users can rename a tree node or click on it to go back to the previous state; allowing them to explore different query possibilities easily and iteratively. Hovering over a node shows a tooltip with the query panel associated with the selected query state. It let users rapidly see what query is associated with each node of the tree If a new change is made on the query from a previous state, a new branch is created on the tree, allowing to revisit and refine explorations. Figure 4.8 shows the provenance tree made to answer question 2, split into 2 branches, with the tooltip showing one of the node query state.

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

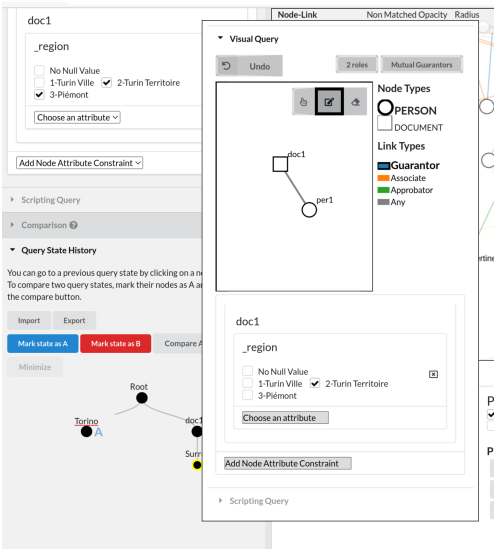


Figure 4.8 – Provenance tree to answer question 2 of collaboration #1: left branch leads to Torino documents (the node is labeled as A) while right branch leads to surrounding documents (the node is labeled as B). The user hovers over one node, revealing a tooltip that shows the visualization of the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

"Compare State A and B" compares them. The interface changes to *comparison mode*. Several buttons appear on top of the provenance tree: A, B, $A-B$, $B-A$, $A \cap B$, and $A \cup B$ for exploring the combinations of the two results of A and B in the two visualizations panels.

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

In the question 6 of collaboration #1, we want to compare the constructions in Torino with the ones in Torino surrounding. Since we previously constructed the query returning all the contracts from Toriono (*Turin*) with the mentioned people, we can return to this point in the provenance tree, and change the constraint of the *region* attribute from *Turin* to *Turin Territoire* and *Piemont* using the checkboxes to get the documents of Torino surroundings in a second query. Both queries are shown in Figure 4.4 (b). The user can then rename the provenance tree nodes with explicit names such as "Torino" and "Surroundings", and mark them as A and B using the appropriate buttons. Clicking on the "Compare State A and B" will make the interface compare the two query results.

Topological Comparison In comparison 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 allows comparing the different graph measures of the two subgraphs by showing them side by side (T3.3). Figure 4.9 shows the comparison table for the queries returning the subgraph of Torino (A) and Torino surroundings (B). Comparing these measures, such as the number of matched documents or the densities, is crucial for SNA. For example, the table indicates that the density is two times higher for Torino, suggesting that less persons paticipate in the same construction compared to Torino surroundings.

Metric	A	B
Bipartite_Clustering_Coefficient	0.52	0.42
Bipartite_Density	0.04	0.02
Bipartite_Redundancy	0.45	0.46
Number_of_Components	13	22
Number_of_DOCUMENT	42	97
Number_of_links	153	419
Number_of_PERSON	99	214
Spectral_Bipartivity	1	1

Showing 1 to 8 of 8 entries Previous 1 Next

Figure 4.9 – Comparison table of the networks measures for Torino subgraph (A) and Torino surroundings subgraph (B).

Attribute-Based Comparison The comparison of one or several attribute distributions between (A) and (B) is also useful for answering the historical questions of our users. In the attribute view (V5) of the results panel, hovering or clicking on an attribute name will show the distribution of this attribute in four contexts: the nodes of the whole graph, the queries (A), (B), and the currently selected Boolean operator (e.g., intersection or union) if one is selected. This allows users to compare attribute distributions between several subsets of interest (T3.2). For example, we can compare the attributes between the contracts of Torino and the ones of its surroundings. We can also compare the persons who worked in Torino, in Torino's close territory, and in both areas, by selecting the intersection operator. Figure 4.10 illustrates the comparison plots for different attributes. The first plot indicates that the types of construction sites differ between the two regions: the city of Torino clearly has a lot of military sites compared to the surroundings of Torino, which has almost none. This is the opposite for the number of religious sites, which are almost all localized in the surroundings of Torino. If we now look at the year distribution of the contracts, we can see a difference in the distributions. The years of Torinos's construction contracts were steady between 1711 and 1717 with a little spike in 1713, while the constructions were more scarce in the surroundings before 1716. We can see a big spike in construction in 1717. This is interesting to our users, as it shows the dynamic of the construction in the area:

the center of the city started to be constructed before other constructions arose in the surroundings.

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

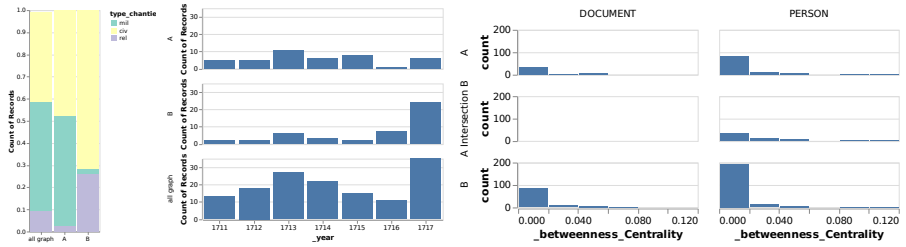


Figure 4.10 – Distribution of the type of constructions, the years, and the betweenness centrality for the documents and signatories of Torino (A), Torino surroundings (B), and the whole graph (top).

4.4.4 Implementation

ComBiNet is made of three components: a web visual interface, a python server, and a Neo4j [100] graph database instance. The client interface is written in JavaScript using D3 [13], Vega [125], and the Trrack library [29]. The python server is written in Flask and interacts with the Neo4j instance for query processing before sending the results to the frontend. We implemented the Cypher parser with the ANTLR parser generator [107]. The abstract syntax tree of the Cypher query is used as a representation of the query. Modifying the query visually update the tree, which is translated into Cypher in the textual query panel. Similarly, a manual change in the Cypher query update the abstract syntax tree which is translated into a visual query.

4.5 Use Cases

In this section, we describe how our system has been able to specifically answer questions from two of our collaborations. the tool was mostly operated by the developers working side by side with the collaborators to test the expressiveness

of the queries and the value of the results visualizations. The tool was refined as needed along the way.

4.5.1 Construction sites in Piedmont (#1)

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

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

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

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

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

4.5.2 French Genealogy (#2)

We describe how ComBiNet allowed us to answer an important question of the use case #2: to detect the largest migrations across several generations, in which areas, and at what time they occurred (question 7 in Table 4.1). The map view

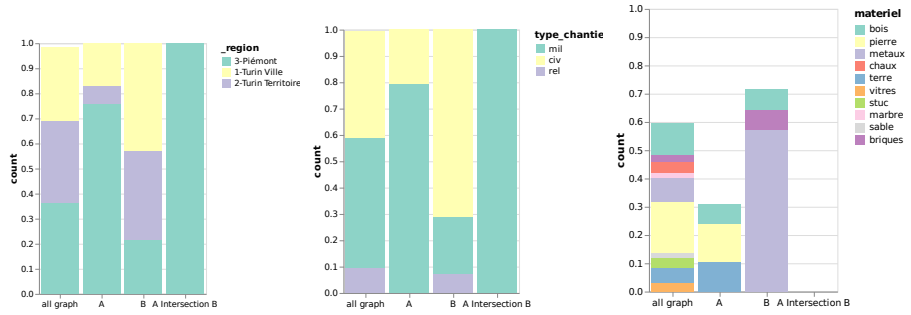
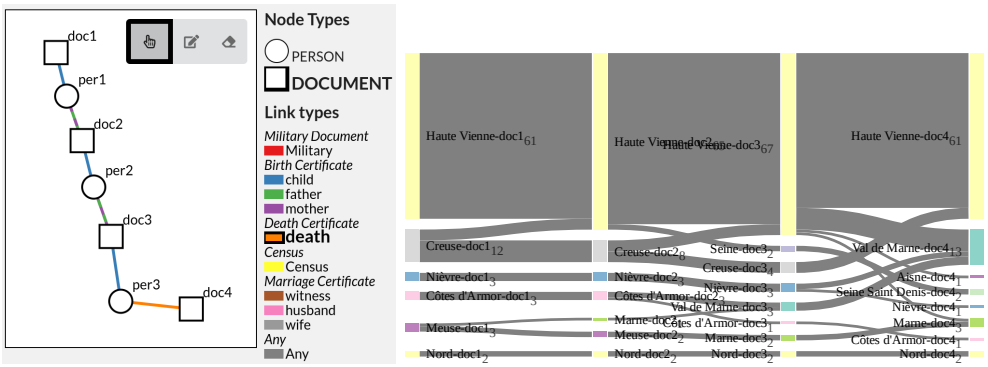


Figure 4.11 – Attributes distributions plots between the whole graph, the *Menafoglio* family (A), the *Zo* family (B), and $A \cap B$, for the *region,type_chantier, material type.*

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.

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,



(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

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.

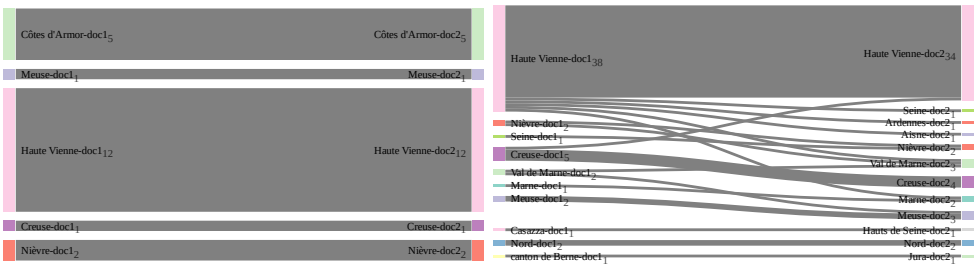


Figure 4.13 – Sankey diagrams showing the migration of people in the 18th and 19th centuries, extracted from their birth and death places.

4.5.3 Marriage acts in Buenos Aires (#3)

I present in this subsection how ComBiNet has been used to detect erroneous encoding during the annotation process of the marriage acts. The documents mention 6659 individuals, whom can have the same name, especially between fathers and sons in this period and region as specified by our collaborator. During the annotation process, the historian and his collaborators gave identifiers to the persons mentioned in the documents—which is the common annotation procedure. However, for this case of homonyms, it can be hard to know if some mentions between different documents refer to the same or different persons. Historians cross the information contained in the different documents to disambiguate the persons [?], but errors can easily be made, i.e., giving the same identifier to different persons or giving different identifiers to the same person. I used ComBiNet in collaboration with researchers of this project to detect erroneous patterns and help them refine their encoded data. For this, we can find the persons mentioned in two acts either

as *husband*, *wife*, or *witness* with a time difference of 70 years or more. Such person nodes in the network are with almost full certainty representing two different persons who lived in different generations. We constructed a request to find this pattern with the visual query view and added the time constraint between the two marriage acts with the Cypher textual input. Figure 4.14 shows the visual query constructed (left), the bipartite view with the persons and documents matching the query highlighted (middle), and the table listing all the documents having mentions of people with erroneous identifiers (bottom right). The table permit to browse through all persons nodes (29 have been found) who correspond in fact to more than one person and to the documents which contain the wrongly given identifiers. Using the exporting capabilities, our collaborator have exported the occurrence table indicating him with their identifiers the pairs of documents mentioning two different persons who have been given the same identifier. Using this table, he has been able to rapidly correct those errors in his annotation framework.

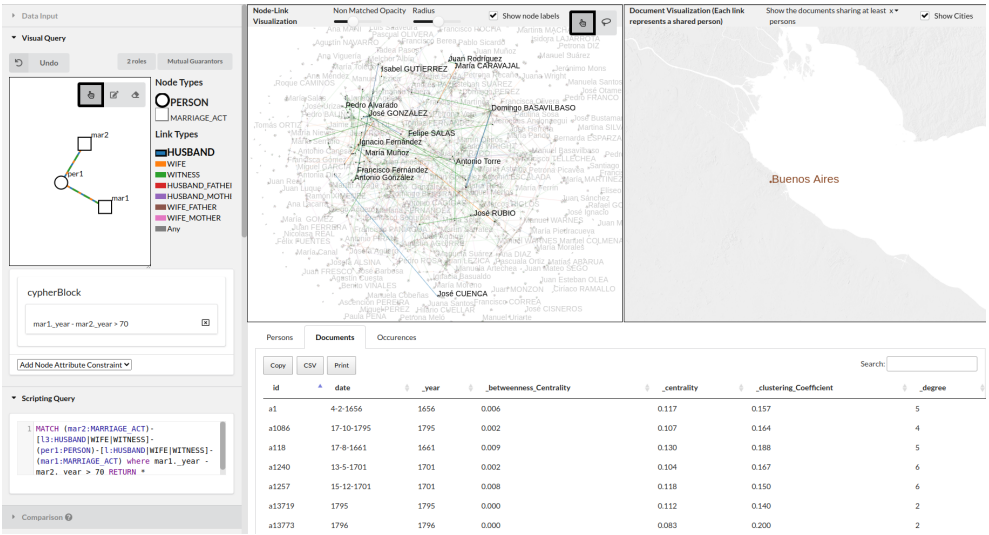


Figure 4.14 – ComBiNet used to request persons appearing as husband, wife, or witness in two marriages which occurred at 70 years apart or more.

4.5.4 Sociology thesis in France

I describe in this third use case how ComBiNet can be used to answer questions about thesis in France between 2016 and 2022. Indeed, some sociological datasets made of documents can also be well modeled as bipartite multivariate dynamic networks like for example thesis dissertations: a thesis is a document with specific attributes such as the subject, the doctoral school, the domain, the university, and the date of defense, and mention several peoples who are socially connected through the thesis defense with different roles: author (*auteur* in french), director(s) (*directeur*), referees (*rapporteur*), and jury president (*président de jury*).

We present here an exploration of the data by ourselves using ComBiNet. A first look at the graph measures tells us that 896 theses have been defended in sociology in France between 2016 and 2021 in France, with 2453 persons included in the defenses (see Figure 4.15 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 theses has been defended all around France, even though the majority are defended in Paris. This is confirmed by a look at the distribution of the cities (Figure 4.15 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 4.15 (right). It means that a lot of thesis defenses include referees and juries from different cities.

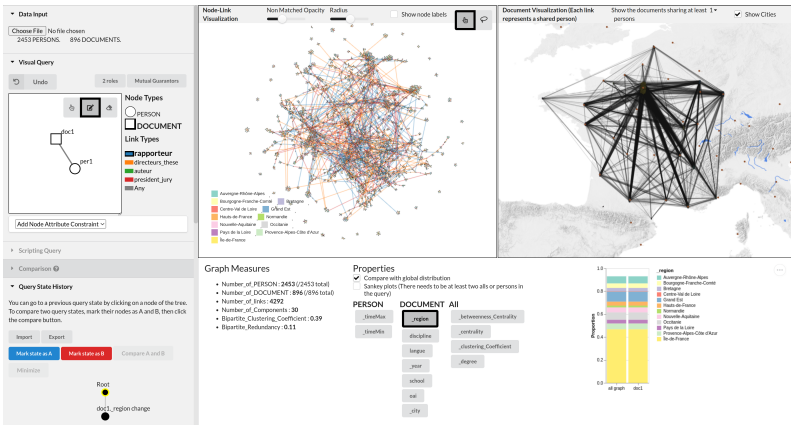


Figure 4.15 – ComBiNet used for exploring theses of sociology defended in France between 2016 and 2021. The bipartite and map views show an overview of two visions of the network. The user selects the *region* attribute, showing the geographical distribution of the defended thesis.

Let's now try to answer an interesting question: "Do referees and jury presidents often ask thesis directors to be referees and jury presidents in their turn of another thesis where they are directors ?". For this, we can construct a visual query representing this pattern by creating two person nodes and two document nodes, and by connecting them with two president links and two referee or jury director links in a symmetrical way, as shown in Figure 4.16 (right). The occurrence table tells us that this pattern has been found 76 times in the network, meaning that this is a recurrent behavior. We are now interested in characterizing the thesis occurring in this pattern, by their regions. We can look at the *city* attribute distribution for this thesis by selecting it in the attribute view as shown in Figure 4.16

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

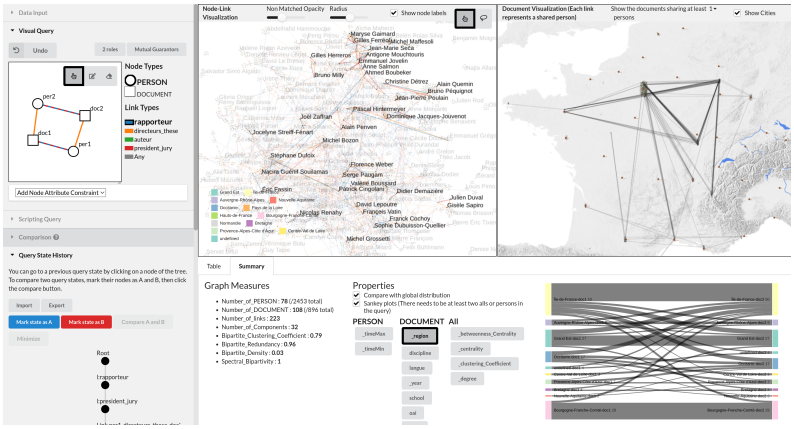


Figure 4.16 – Sociology thesis dataset explored with ComBiNet. The user constructed a visual query to see if there are symmetrical relationships between thesis directors and referees (or jury directors). The *region* attribute is selected with the Sankey option, letting the user see if there are correlations between the regions of the thesis found in this pattern.

4.6 Formative Usability Study

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

document tables (which are updated after each query), the persistent selection of nodes across the two views and the tables, and the undo feature for visual queries. At the final meetings, the three users were able to perform exploration, queries, and comparisons to answer socio-historical questions by themselves.

4.6.1 Feedback

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

4.7 Discussion

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

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

Generalizibility. The system has been designed specifically for bipartite multivariate dynamic networks, which models well a diversity of historical sources we encountered via our collaborations: marriage acts, birth/death certificates, construction/work contracts, census, and migrations forms. Moreover, bipartite multivariate dynamic network can also be used to model other similar data types, such as scientific publications or thesis data. However, other kinds of historical textual data exist where documents can mention each other, such as in private letters for example. The model and interface would need to be slightly modified to take

into account document-to-document links for these datasets. Bipartite networks are also used in various other disciplines, such as biology [73] and chemistry [74]. ComBiNet could be extended to these other application domains, in particular by modifying the map view to show other location data related to the entities of the network, or removing it altogether if it makes no sense for a particular domain.

4.8 Conclusion and Future Work

I presented in this chapter ComBiNet, a VA system for exploring social networks modeled from historical textual sources, primarily aimed at social historians. 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. ComBiNet relies on this data model to let historians explore a concrete representation of their annotated documents (i.e., the model express the *reality* of the documents, without the use of projections or distortions) with *traceability* to the original sources and *simplicity*. Historians can hence reflect on their encoding process and answer their socio-historical questions using 1) dynamic queries on the network structure and attributes to highlight groups of interests or erroneous patterns 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. I have shown that complex explorations and analyses were easy to perform, and validated our approach by first describing four use cases among many more projects we are collaborating with and by performing a formative usability study showing that after many improvements the system is usable by social scientists.

By specifying a unifying data model and novel high-level visual and interactive mechanisms for comparing topology, attributes, and time, social scientists were able to correct their data more easily with exploration 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 easy-to-use interactions implemented in ComBiNet such as coordinated views, visual querying, and comparison. Using these mechanisms, social scientists performed visual exploratory analyses of their network based on topological and attribute descriptions, and comparisons of subgroups of interests—between them or the overall network. This methodology allows them to either ground or refute their hypotheses in network measures and attribute distributions, or to generate new ones from new insight revealed thanks to the 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 *reality* of the

documents, *traceability* of the network and results, and *simplicity* of use, which are essential for historical work.

For future work, ComBiNet could be extended to support more SNA measures and computations such as clustering; it would create a new attribute containing a cluster identifier. The interface currently proposes two layouts based on the topology and the geolocations of the entities. Providing more layout options could be interesting, especially one to highlight better the time, similar to the PAOHvis technique [142]. Finally, the interface could in the future make suggestions on the query construction process based on frequent subgraphs similar to VERTIGo [28], and within a mixed-initiative perspective [**?**].

In the next chapter, I present a visual interface following this mixed-initiative framework for network clustering, to answer **Q3** and showing that such approaches can help social scientists use data mining tools while controlling their biases—with the condition of an explicit results' traceability.

5 PK-Clustering

This chapter is an updated version of an article published in IEEE Transactions on Visualization and Computer Graphics (TVCG) 2020. It was a joint work with my collaborators Paolo Buono, Catherine Plaisant, Jean-Daniel Fekete, and Paola Valdivia. I participated in the development of the interface, discussions, evaluation, and writing of the paper.

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 [114], Gephi [5], Python with NetworkX [57], or Pajek [106]. 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 re-

sults 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 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.

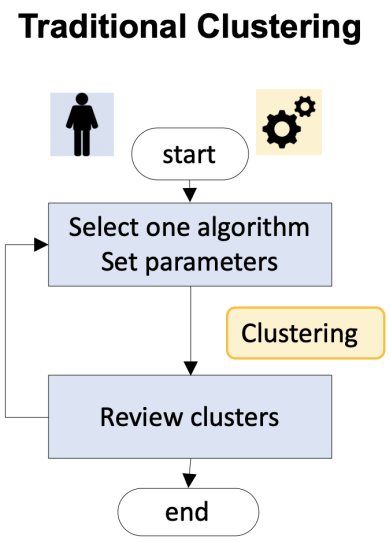


Figure 5.1 – Traditional Clustering. The output is a clustering, usually from a randomly chosen algorithm.

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 com-

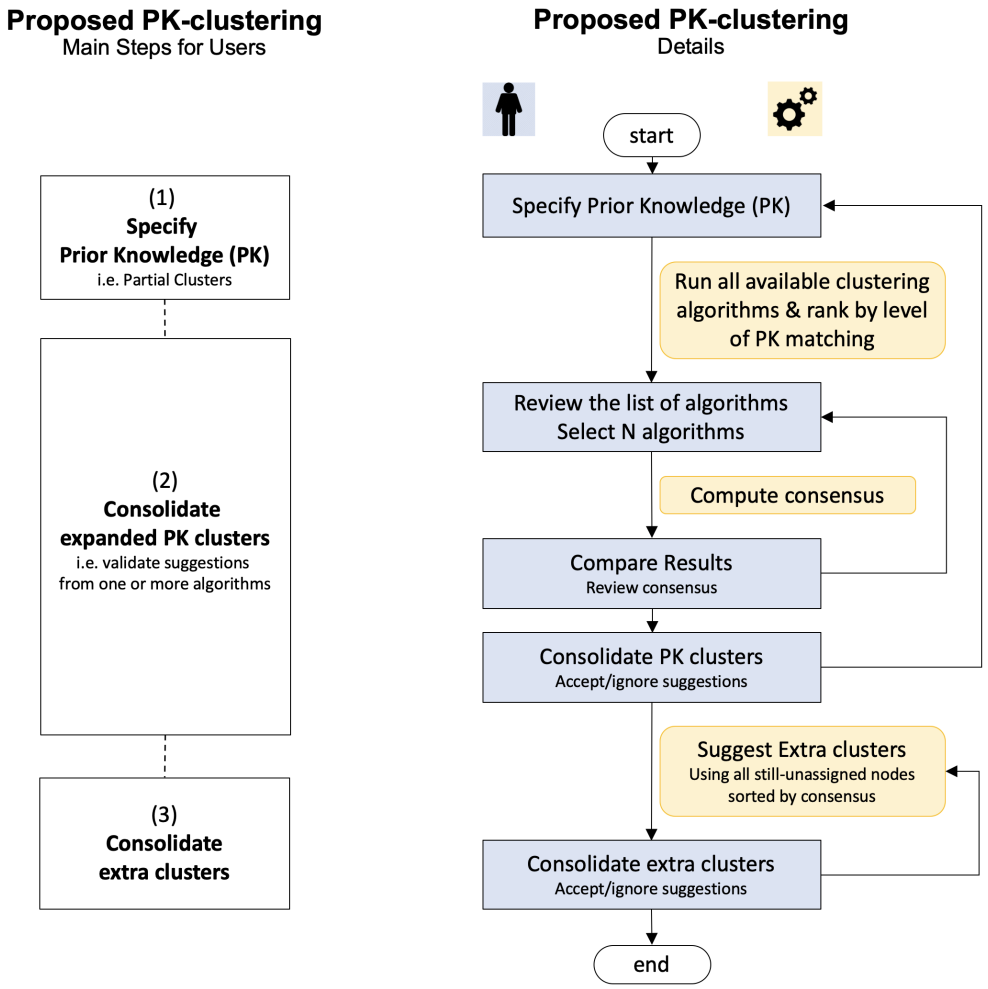


Figure 5.2 – PK-clustering. The output is a clustering supported by algorithms and validated (fully or partially) according to the user's Prior Knowledge.

pare 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 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 §5.3.1).

Contrary to a current trend [92], 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 multidimensional 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 [119] 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 [6, 147]. 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 [11], Density-Based Clustering [?], and graph-based clustering [147]. 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 [126] 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. [?, ?, ?, 151], 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. [145]. 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 [124], using new tasks assessing group-level understanding.

Holten [?] proposes edge bundling on compound graphs. He bundles together adjacenct 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 out-

puts 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 [137]. 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 §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 being 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 encounter. 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.

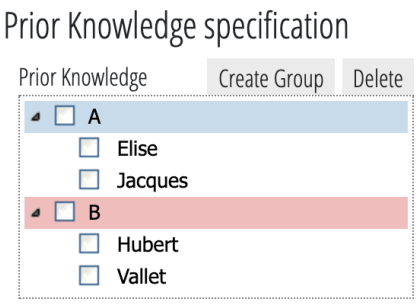


Figure 5.3 – Prior Knowledge specification, the user defined two groups composed of two members.

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 §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 $N = 1, 2, ..., 12$, 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.

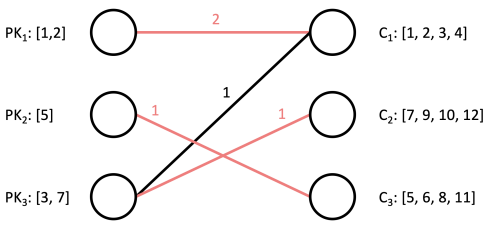


Figure 5.4 – Red edges represent the prior knowledge matching

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 propagation/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 rep-

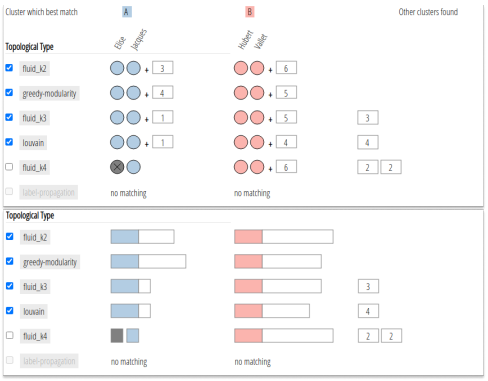


Figure 5.5 – Two different modalities for the ranked list of algorithms. Top: persons are shown as circles. Bottom: aggregated view. Colors indicate the matching group. Gray indicates no match. White indicates extra nodes or clusters.

resented 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 matching, 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

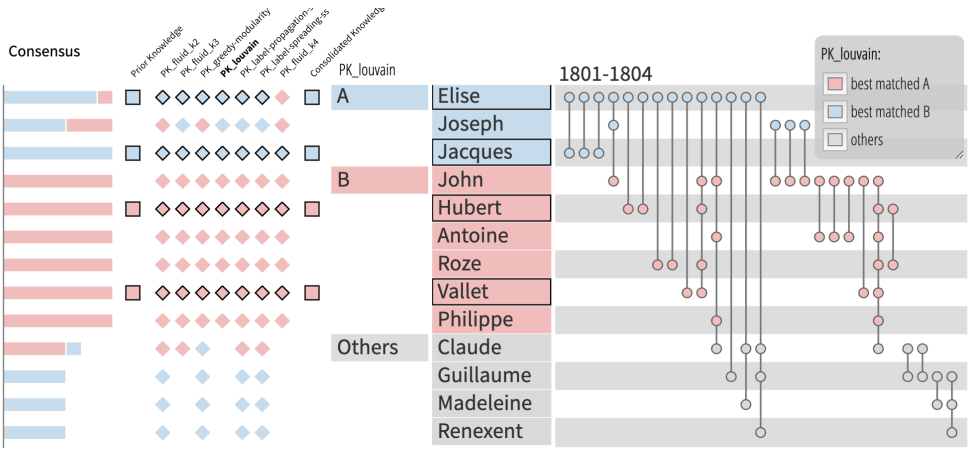


Figure 5.6 – Reviewing and comparing results of multiple algorithms. One algorithm is selected to order the names and group them, but icons show how other algorithms cluster the nodes differently, summarized in the consensus bar on the left.

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 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,

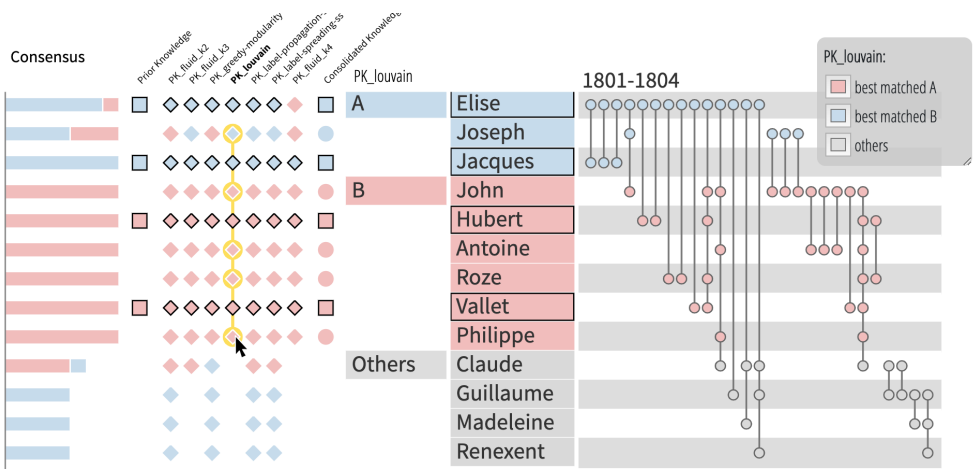


Figure 5.7 – The user quickly drags on consecutive icons (in yellow) representing the suggestions made by one algorithm to validate node clustering. Once the cursor is released the validated nodes appear as squares icons in the Consolidated Knowledge column.

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. iThe 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 §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.

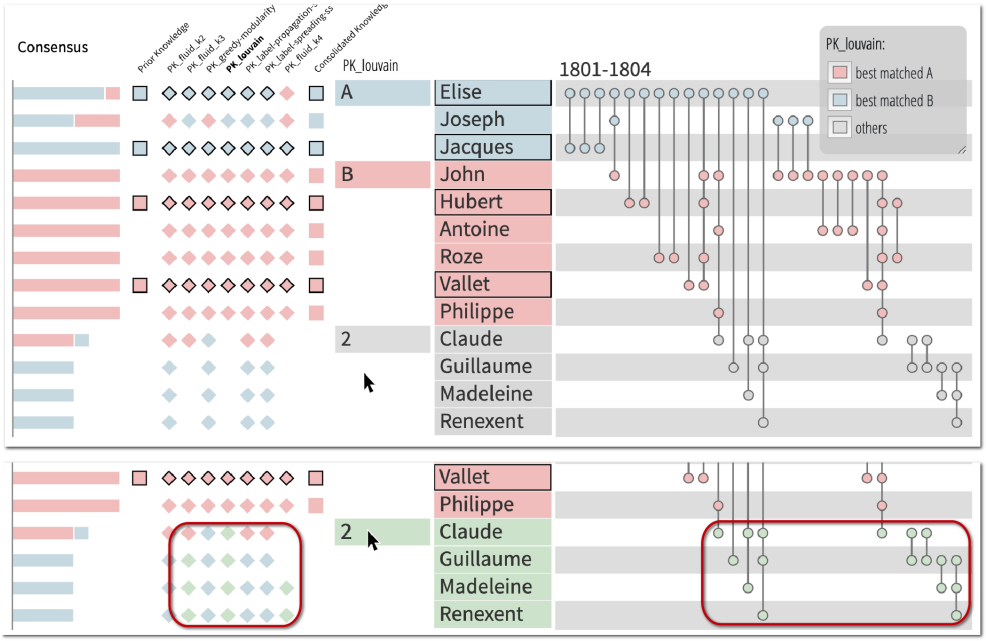


Figure 5.8 – Suggestion of extra clusters. The two PK-groups (red and blue) are validated (nodes in the consensus column are all squared). One extra clusters is proposed by the Louvain algorithm, labeled as 2. Hovering over the cluster 2, the consensus is displayed by the green diamonds. This feedback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

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 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 (§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.

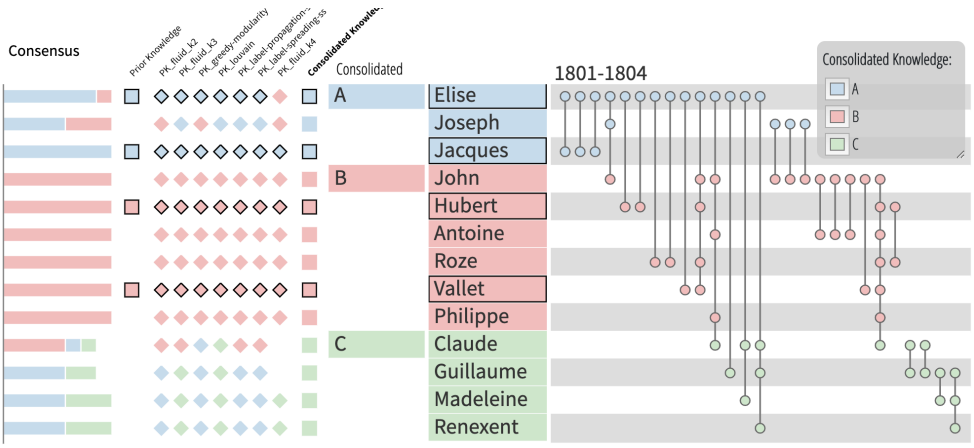


Figure 5.9 – The dataset has been fully consolidated. The persons are grouped and colored by the consolidated knowledge. The user decided to assign Claude, Guillaume, Madeleine and Renexent to cluster C, by taking into account the graph and the consensus of the algorithms.

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 comfident 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.

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

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 [67] of: • academic lineages • research topics • job (academia, industry) • geography (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

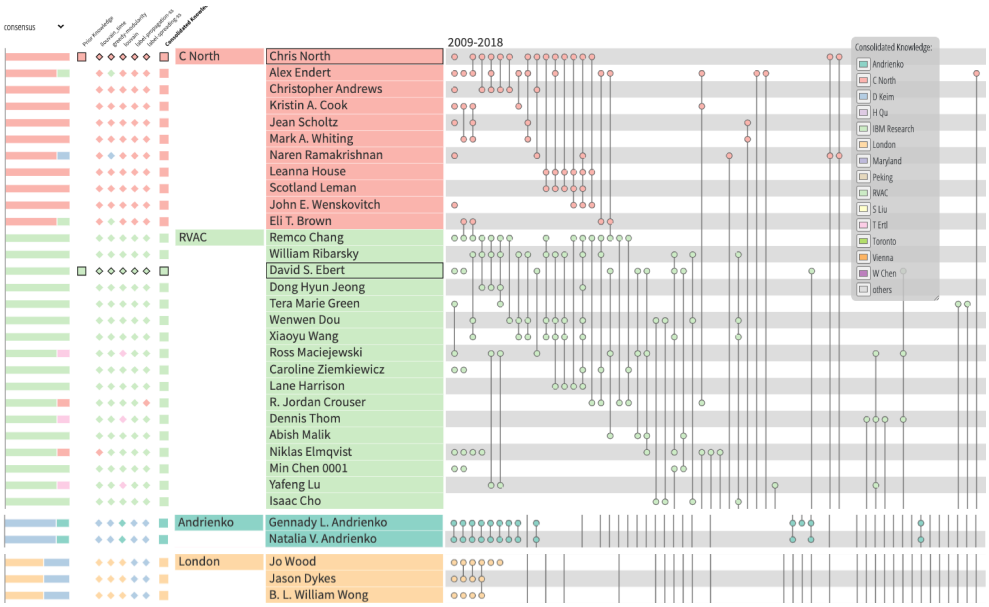


Figure 5.11 – Four consolidated groups in the VAST dataset: C North, RVAC, Andrienko and London

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 [146], 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 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 §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 $\times$ 8 and 16 Gigabytes of memory. For the full Marie Boucher social network described in §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 §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 [?, 116] 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 lifes 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 bipartite multivariate dynamic networks 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 bipartite multivariate dynamic networks, 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, specifics 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 od 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 descibred 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 [81].

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 worlflow 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 [136]. 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 definying 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 §6.2, temporality is one of the key dimension of historical networks modeleds as bipartite multivariate dynamic

networks. Several layouts have been proposed to show the dynamic aspect of dynamic graphs and bipartite dynamic graph, such as PAOHVIS [142], which is the layout PK-Clustering is based on. However, this type of layout do not allow to see the attributes of persons and documents, nor the geolocation of entities. Moreover, most proposed layouts do not scale beyond approximately one hundred nodes. One way to solve this scalability problem is to aggregate the network, for example using dynamic clustering. Several dynamic clustering algorithms have been developed for dynamic networks, but they often struggle to take into account complex dynamic of clusters, such as merging, splitting, or community drifting. Furthermore, no layout currently exist to visually display those dynamic communities specifically for dynamic hypergraphs. More work can still be done in this space, to visualize the dynamic aspect of bipartite multivariate dynamic networks, and visualizing dynamic communities. **talk about prototype** As with static clustering, the results produced by algorithms can be biased, and several interesting dynamic partitions coexist. A similar procedure of PK-Clustering but with dynamic clustering would mitigate those problems, by providing a way to construct meaningful clusters for social scientists using their prior-knowledge, algorithmic consensus, and exploration capabilities.

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 [110], or link prediction [90]. 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 [62]. 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.

combining both projects

Taking into account uncertainty of sources in the network model

6.4 Conclusion

Bibliography

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

[2] Mashael Alkadi, Vanessa Serrano, James Scott-Brown, Catherine Plaisant, Jean-Daniel Fekete, Uta Hinrichs, and Benjamin Bach. Understanding barriers to network exploration with visualization: A report from the trenches. *IEEE Transactions on Visualization and Computer Graphics*, 27(2), 2022. 12, 16, 17, 19, 35, 39, 41, 42, 44, 47, 56

[3] Keith Andrews, Martin Wohlfahrt, and Gerhard Wurzinger. Visual Graph Comparison. In *2009 13th International Conference Information Visualisation*, pages 62–67, July 2009. 63

[4] F. J. Anscombe. Graphs in Statistical Analysis. *The American Statistician*, 27(1):17–21, February 1973. 20, 22

[5] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. In Eytan Adar, Matthew Hurst, Tim Finin, Natalie S. Glance, Nicolas Nicolov, and Belle L. Tseng, editors, *ICWSM*. The AAAI Press, 2009. 15, 38, 60, 69, 91

[6] Sugato Basu, Ian Davidson, and Kiri Wagstaff. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC, first edition, 2008. 95

[7] Leilani Battle and Jeffrey Heer. Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau. *Computer Graphics Forum*, 38(3):145–159, 2019. 63

[8] Jacques Bertin. *Sémiologie graphique: les diagrammes, les réseaux, les cartes*. Paris: Gauthier-Villars, 1967. 20, 21

[9] A. Bezerianos, F. Chevalier, P. Dragicevic, N. Elmqvist, and J.d. Fekete. GraphDice: A System for Exploring Multivariate Social Networks. *Computer Graphics Forum*, 29(3):863–872, 2010. 70

[10] Marc Bloch. *Apologie Pour l'histoire*. A. Colin, 1949. 12

[11] Christian Böhm and Claudia Plant. HISSCLU: A hierarchical density-based method for semi-supervised clustering. In *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '08, pages 440–451, New York, NY, USA, 2008. ACM. 95

[12] Christian Bors, John Wenskovitch, Michelle Dowling, Simon Attfield, Leilani Battle, Alex Endert, Olga Kulyk, and Robert S. Laramee. A Provenance Task Abstraction Framework. *IEEE Computer Graphics and Applications*, 39(6):46–60, November 2019. 63

[13] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, December 2011. 69, 80

[14] Pierre Bourdieu. Sur les rapports entre la sociologie et l'histoire en Allemagne et en France. *Actes de la Recherche en Sciences Sociales*, 106(1):108–122, 1995. 25

[15] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On Modularity Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188, February 2008. 47

[16] Peter Burke. *History and Social Theory*. Polity, 2005. 25

[17] Steven P. Callahan, Juliana Freire, Emanuele Santos, Carlos E. Scheidegger, Cláudio T. Silva, and Huy T. Vo. VisTrails: Visualization meets data management. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data - SIGMOD '06*, page 745, Chicago, IL, USA, 2006. ACM Press. 63

[18] Charles-Olivier Carbonell. *L'Historiographie*. FeniXX, January 1981. 25

[19] Stuart-K. Card, Jock-D. Mackinlay, and Ben Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers In, San Francisco, Calif, February 1999. 14, 20

[20] Duen Horng Chau, Christos Faloutsos, Hanghang Tong, Jason I. Hong, Brian Gallagher, and Tina Eliassi-Rad. GRAPHITE: A Visual Query System for Large Graphs. In *2008 IEEE International Conference on Data Mining Workshops*, pages 963–966, December 2008. 62

[21] J. S. Coleman. Introduction to mathematical sociology. *Introduction to mathematical sociology.*, 1964. 31

[22] TEI Consortium. TEI P5: Guidelines for electronic text encoding and interchange, February 2021. 46

[23] Pascal Cristofoli. Aux sources des grands réseaux d'interactions. *Reseaux*, 152(6):21–58, 2008. 11, 17, 33, 42, 44, 47, 53, 60

[24] Pascal Cristofoli. Principes et usages des dessins de réseaux en SHS. *La visualisation des données en histoire*, page 35, 2015. 12, 37, 69

[25] Pascal Cristofoli and Nicoletta Rolla. Temporalités à l'œuvre dans les chantiers du bâtiment. *Temporalités. Revue de sciences sociales et humaines*, (27), June 2018. 49, 52, 64, 69

[26] Tarik Crnovrsanin, Chris W. Muelder, Robert Faris, Diane Felmlee, and Kwan-Liu Ma. Visualization techniques for categorical analysis of social networks with multiple edge sets. *Social Networks*, 37:56–64, 2014. 49

[27] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. 62, 69

[28] Erick Cuenca, Arnaud Sallaberry, Dino Ienco, and Pascal Poncelet. VERTIGo: A Visual Platform for Querying and Exploring Large Multilayer Networks. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2021. 62

[29] Zach Cutler, Kiran Gadhave, and Alexander Lex. Trrack: A Library for Provenance-Tracking in Web-Based Visualizations. In *2020 IEEE Visualization Conference (VIS)*, pages 116–120, October 2020. 77, 80

[30] Mandeep K. Dhami, Ian K. Belton, and David R. Mandel. The "analysis of competing hypotheses" in intelligence analysis. *Applied Cognitive Psychology*, 33(6):1080–1090, 2019. 14

[31] Jana Diesner, Craig Evans, and Jinseok Kim. Impact of Entity Disambiguation Errors on Social Network Properties. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):81–90, 2015. 40, 42, 46

[32] Dana Diminescu. The migration of ethnic germans from romania to west germany: Insights from the archives of the former communist regime. In *CERS, Public Lecture, UCLA*, Los Angeles, United States, March 2020. 52, 66

[33] Nicole Dufournaud. La recherche empirique en histoire à l'ère numérique. *Gazette des archives*, 240(4):397–407, 2015. 11

[34] Nicole Dufournaud. Comment rendre visible le rôle économique des femmes sous l'Ancien Régime ? Étude méthodologique sur les marchandes à Nantes aux XVIe et XVIIe siècles. In Bernard Michon and Nicole Dufournaud, editors, *Femmes et Négoce Dans Les Ports Européens (Fin Du Moyen Age - XIXe Siècle)*, pages 65–84. Peter Lang, 2018. 16, 44, 48

[35] Nicole Dufournaud and Jean-Daniel Fekete. Comparaison d'outils pour la visualisation de sources historiques codées en XML/TEI. *Document numérique*, 9(2):37–56, April 2006. 46

[36] Cody Dunne, Nathalie Henry Riche, Bongshin Lee, Ronald Metoyer, and George Robertson. GraphTrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1663–1672, New York, NY, USA, May 2012. Association for Computing Machinery. 63

[37] P. Erdös and A. Rényi. On the evolution of random graphs. In *On the Evolution of Random Graphs*, pages 38–82. Princeton University Press, October 2011. 30

[38] Emily Erikson and Peter Bearman. Malfeasance and the Foundations for Global Trade: The Structure of English Trade in the East Indies, 1601–1833. *American Journal of Sociology*, 112(1):195–230, July 2006. 49

[39] Michael Eve. Deux traditions d'analyse des reseaux sociaux. *Réseaux*, 115(5):183–212, 2002. 17, 32, 33

[40] Lucien Febvre. VERS UNE AUTRE HISTOIRE. *Revue de Métaphysique et de Morale*, 54(3/4):225–247, 1949. 25

[41] Jean-Daniel Fekete, Danyel Fisher, Arnab Nandi, and Michael Sedlmair. *Progressive Data Analysis and Visualization*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, April 2019. 75

[42] L. Freeman. Visualizing Social Networks. *J. Soc. Struct.*, 2000. 12, 14, 36

[43] L.C. Freeman. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press, 2004. 11, 13, 19, 29, 30, 31, 32, 47

[44] Manuel Freire, Catherine Plaisant, Ben Shneiderman, and Jen Golbeck. ManyNets: An interface for multiple network analysis and visualization. In *CHI '10*, CHI '10, pages 213–222, New York, NY, USA, 2010. ACM. 63

[45] Michael Friendly. Visions and Re-Visions of Charles Joseph Minard. *Journal of Educational and Behavioral Statistics*, 27(1):31–51, March 2002. 20

[46] Michael Friendly. A Brief History of Data Visualization. In Chun-houh Chen, Wolfgang Härdle, and Antony Unwin, editors, *Handbook of Data Visualization*, Springer Handbooks Comp.Statistics, pages 15–56. Springer, Berlin, Heidelberg, 2008. 20

[47] GEDCOM: The genealogy data standard. 35

[48] Mohammad Ghoniem, J.-D. Fekete, and Philippe Castagliola. A comparison of the readability of graphs using node-link and matrix-based representations.

In *IEEE Symposium on Information Visualization*, pages 17–24. Ieee, 2004. 37

[49] Carlo Ginzburg and Carlo Poni. La micro-histoire. *Le Débat*, 17(10):133, 1981. 13, 33, 43

[50] Barney G. Glaser and Anselm L. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Transaction, New Brunswick, 5. paperback print edition, 2010. 14

[51] Michael Gleicher. Considerations for visualizing comparison. *IEEE Trans. Vis. Comput. Graphics*, 24(1):413–423, 2018. 63

[52] Claudia Goldin. Cliometrics and the Nobel. *Journal of Economic Perspectives*, 9(2):191–208, June 1995. 13

[53] Martin Grandjean. Social network analysis and visualization: Moreno's Sociograms revisited, 2015. 30

[54] Martin Grandjean. Analisi e visualizzazioni delle reti in storia. L'esempio della cooperazione intellettuale della Società delle Nazioni. *ME*, (2/2017), 2017. 60

[55] Maurizio Gribaudi and Alain Blum. Des catégories aux liens individuels : l'analyse statistique de l'espace social. *Annales*, 45(6):1365–1402, 1990. 12

[56] Jo Guldi and David Armitage. *The History Manifesto*. Cambridge University Press, October 2014. 11

[57] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11–15, Pasadena, CA USA, 2008. 62, 91

[58] Klaus Hamberger, Cyril Grange, Michael Houseman, and Christian Momon. Scanning for patterns of relationship: Analyzing kinship and marriage networks with Puck 2.0. *The History of the Family*, 19(4):564–596, October 2014. 33, 36, 49

[59] Klaus Hamberger, Michael Houseman, and R. White, Douglas. Kinship network analysis. In John Scott & Peter J. Carrington, editor, *The Sage Handbook of Social Network Analysis*, pages 533–549. Sage Publications, 2011. 11, 35

[60] Mountaz Hascoët and Pierre Dragicevic. Interactive graph matching and visual comparison of graphs and clustered graphs. In Genny Tortora, Stefano Levialdi, and Maurizio Tucci, editors, *AVI '12*, pages 522–529. ACM, 2012. 63

[61] J. Heer and D. Boyd. Vizster: Visualizing online social networks. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 32–39, October 2005. 71

[62] Jeffrey Heer. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, 116(6):1844–1850, 2019. 120

[63] Louis Henry and Michel Fleury. Des registres paroissiaux a l'histoire de la population: Manuel de dépouillement et d'exploitation de l'état civil ancien. *Population (French Edition)*, 11(1):142–144, 1956. 13

[64] Nathalie Henry, Jean-Daniel Fekete, and Michael J. McGuffin. NodeTrix: A Hybrid Visualization of Social Networks. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1302–1309, November 2007. 37, 38

[65] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, and Sabrina Kirrane et al. Knowledge graphs. *ACM Comput. Surv.*, 54(4), July 2021. 49

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

[67] Infovis SC policies FAQ. 110

[68] Piers J Ingram, Michael PH Stumpf, and Jaroslav Stark. Network motifs: Structure does not determine function. *BMC Genomics*, 7:108, May 2006. 62

[69] Frédéric Kaplan. The Venice Time Machine. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, DocEng '15, page 73, New York, NY, USA, September 2015. Association for Computing Machinery. 16, 27

[70] Karine Karila-Cohen, Claire Lemercier, Isabelle Rosé, and Claire Zalc. Nouvelles cuisines de l'histoire quantitative. *Annales. Histoire, Sciences Sociales*, 73(4):773–783, December 2018. 11, 43, 44, 48

[71] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual Analytics: Definition, Process, and Challenges. In Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North, editors, *Information Visualization: Human-Centered Issues and Perspectives*, Lecture Notes in Computer Science, pages 154–175. Springer, Berlin, Heidelberg, 2008. 15, 16, 23

[72] Florian Kerschbaumer, Linda von Keyserlingk-Rehbein, Martin Stark, and Marten Düring. *The Power of Networks. Prospects of Historical Network Research*. Routledge, December 2021. 11, 19, 34

[73] Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and cellular networks. *PLoS computational biology*, 5(5):e1000385, 2009. 88

[74] Elena V. Konstantinova and Vladimir A. Skorobogatov. Application of hypergraph theory in chemistry. *Discrete Mathematics*, 235(1-3):365–383, May 2001. 88

[75] C. Kosak, J. Marks, and S. Shieber. Automating the layout of network diagrams with specified visual organization. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(3):440–454, March 1994. 37

[76] Ernest Labrousse. *La Crise de l'économie Française à La Fin de l'Ancien Régime et Au Début de La Révolution*, volume 1. Presses Universitaires de France-PUF, 1990. 25

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

[78] Katherine A. Larson. Thomas F. Tartaron, Maritime Networks in the Mycenaean World. New York: Cambridge University Press, 2013. *Comparative Studies in Society and History*, 56(4):1064–1065, October 2014. 11

[79] Emmanuel Lazega. *Réseaux sociaux et structures relationnelles*. Presses universitaires de France, Paris, 1998. 14, 32

[80] Claire Lemercier. 12. Formal network methods in history: Why and how? In Georg Fertig, editor, *Social Networks, Political Institutions, and Rural Societies*, volume 11, pages 281–310. Brepols Publishers, Turnhout, January 2015. 11, 16, 17, 19, 31, 33, 34, 35, 36, 39, 41, 50, 60

[81] Claire Lemercier and Claire Zalc. *Quantitative Methods in the Humanities: An Introduction*. University of Virginia Press, March 2019. 12, 14, 26, 27, 29, 36, 41, 43, 44, 60, 119

[82] Claire Lemercier and Claire Zalc. Back to the Sources: Practicing and Teaching Quantitative History in the 2020s. *Capitalism*, 2(2):473–508, 2021. 16, 26, 41, 42, 43, 44, 48

[83] Bernard Lepetit. L'histoire quantitative : deux ou trois choses que je sais d'elle. *Histoire & Mesure*, 4(3):191–199, 1989. 11, 43

[84] Carola Lipp. Kinship Networks, Local Government, and Elections in a Town in Southwest Germany, 1800-1850. *Journal of Family History*, 30(4):347–365, October 2005. 33

[85] Gribaudi Maurizio. *Espaces, Temporalités, Stratifications :. Exercices Méthodologiques Sur Les Réseaux Sociaux*. Editions de l'Ecole des Hautes Etudes en Sciences Sociales, Paris, January 2000. 32

[86] Philip Mayer. Migrancy and the Study of Africans in Towns. *American Anthropologist*, 64(3):576–592, 1962. 33

[87] Fintan McGee, Benjamin Renoust, Daniel Archambault, Mohammad Ghoniem, Andreas Kerren, and Bruno Pinaud et al. *Visual Analysis of Multilayer Networks*. Synthesis Lectures on Visualization. Morgan & Claypool Publishers, 2021. 49

[88] Michael J. McGuffin. Simple algorithms for network visualization: A tutorial. *Tsinghua Science and Technology*, 17(4):383–398, August 2012. 37

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

[90] R. Michalski, P. Kazienko, and D. Krol. Predicting Social Network Measures Using Machine Learning Approach. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1056–1059, Istanbul, August 2012. IEEE. 120

[91] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594):824–827, October 2002. 32, 62

[92] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. Lulu.com, 2019. 94

[93] Megan Monroe, Krist Wongsuphasawat, Catherine Plaisant, Ben Shneiderman, Jeff Millstein, and Sigfried Gold. Exploring point and interval event patterns: Display methods and interactive visual query. *University of Maryland Technical Report*, 2012. 87

[94] J. L. Moreno. *Who Shall Survive?: A New Approach to the Problem of Human Interrelations*. Who Shall Survive?: A New Approach to the Problem of Human Interrelations. Nervous and Mental Disease Publishing Co, Washington, DC, US, 1934. 30, 36

[95] J. L. Moreno. Foundations of Sociometry: An Introduction. *Sociometry*, 4(1):15, February 1941. 30, 49

[96] Zacarias Moutoukias. Buenos Aires, port between two oceans: Mobilities, networks, stratifications (2nd half of the 18th century). *E-SPANIA-REVUE ELECTRONIQUE D ETUDES HISPANIQUES MEDIEVALES*, 25, 2016. 52, 66

[97] Zacharias Moutoukias. Réseaux personnels et autorité coloniale : Les négociants de Buenos Aires au XVIIIe siècle. *Annales. Histoire, Sciences Sociales*, 47(4-5):889–915, October 1992. 33

[98] Andrej Mrvar and Vladimir Batagelj. Analysis and visualization of large networks with program package Pajek. *Complex Adaptive Systems Modeling*, 4(1), April 2016. 38, 69

[99] Natural earth. 70

[100] Neo4j graph data platform. 61, 62, 80

[101] Rolla Nicoletta. Mobilité et conflits. Travailler sur les chantiers de construction piémontais dans la première moitié du XVIIIe siècle. In Andrea Caracausi and Marco Schnyder, editors, *Travail et Mobilité En Europe (XVIe-XIXe Siècles)*, Coll. Histoire et Civilisations. Presses universitaires du Septentrion, Villeneuve d'Ascq, 2018. 52

[102] Carolina Nobre, Marc Streit, and Alexander Lex. Juniper: A Tree+Table Approach to Multivariate Graph Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):544–554, January 2019. 37

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

[104] Maryjane Osa. *Solidarity And Contention: Networks Of Polish Opposition*. Univ Of Minnesota Press, Minneapolis, first edition edition, July 2003. 11, 47

[105] John F. Padgett and Christopher K. Ansell. Robust Action and the Rise of the Medici, 1400-1434. *American Journal of Sociology*, 98(6):1259–1319, May 1993. 11, 12, 15, 34

[106] Pajek — Analysis and visualization of very large networks. 15, 91

[107] Terence J. Parr and Russell W. Quong. ANTLR: A predicated-LL (k) parser generator. *Software: Practice and Experience*, 25(7):789–810, 1995. 80

[108] Vanessa Peña-Araya, Tong Xue, Emmanuel Pietriga, Laurent Amsaleg, and Anastasia Bezerianos. HyperStorylines: Interactively untangling dynamic hypergraphs. *Information Visualization*, 21(1):38–62, January 2022. 53

[109] Cindarella Sarah Maria Petz. *On Combining Network Research and Computational Methods on Historical Research Questions and Its Implications for the Digital Humanities*. PhD thesis, Technische Universität München, 2022. 34, 43

[110] James P. Philips and Nasseh Tabrizi. Historical Document Processing: Historical Document Processing: A Survey of Techniques, Tools, and Trends, September 2020. 120

[111] Robert Pienta, Fred Hohman, Alex Endert, Acar Tamersoy, Kevin Roundy, Chris Gates, Shamkant Navathe, and Duen Horng Chau. VIGOR: Interactive Visual Exploration of Graph Query Results. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):215–225, January 2018. 63

[112] Alexis Pister, Nicole Dufournaud, Pascal Cristofoli, Christophe Prieur, and Jean-Daniel Fekete. From Historical Documents To Social Network Visualization: Potential Pitfalls and Network Modeling. In *7th Workshop on Visualization for the Digital Humanities (VIS4DH)*, 2022. 41

[113] Antoine Prost. *Douze Leçons sur l'histoire*. Média Diffusion, April 2014. 12, 13, 19, 24, 25

[114] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. 62, 91

[115] Eric Ragan, Endert Alex, Jibonananda Sanyal, and Jian Chen. Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes. *IEEE Transactions on Visualization and Computer Graphics*, 22(1), January 2016. 63

[116] Ramana Rao and Stuart K. Card. The table lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '94, pages 318–322, New York, NY, USA, 1994. Association for Computing Machinery. 115

[117] Donghao Ren, Bongshin Lee, and Matthew Brehmer. Charticulator: Interactive Construction of Bespoke Chart Layouts. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):789–799, January 2019. 63

[118] Pedro Ribeiro and Fernando Silva. Discovering Colored Network Motifs. In Pierluigi Contucci, Ronaldo Menezes, Andrea Omicini, and Julia Poncela-Casasnovas, editors, *Complex Networks V*, Studies in Computational Intelligence, pages 107–118, Cham, 2014. Springer International Publishing. 62

[119] Giulio Rossetti and Rémy Cazabet. Community discovery in dynamic networks: A survey. *ACM Comput. Surv.*, 51(2), February 2018. 95

[120] Fabrice Rossi, Nathalie Vialaneix, and Florent Hautefeuille. Exploration of a large database of French notarial acts with social network methods. *Digital Medievalist*, 9:2013, July 2014. 60

[121] Juan A. Rubio-Mondejar and Josean Garrues-Irurzun. Women entrepreneurs and family networks in Andalusia (Spain) during the second industrial revolution. *Business History*, pages 1–22, May 2022. 11

[122] C.J. Rueda and Catedral de Buenos Aires. *Matrimonios de La Catedral de Buenos Aires, 1747-1823*. Number v. 2 in Fuentes Históricas y Genealógicas Argentinas. Fuentes Históricas y Genealógicas Argentinas, 1989. 52

[123] Anni Sairio. Methodological and practical aspects of historical network analysis: A case study of the Bluestocking letters. In Arja Nurmi, Minna Nevala, and Minna Palander-Collin, editors, *Pragmatics & Beyond New Series*, volume 183, pages 107–135. John Benjamins Publishing Company, Amsterdam, 2009. 48

[124] Bahador Saket, Paolo Simonetto, and Stephen Kobourov. Group-level graph visualization taxonomy. In N. Elmqvist, M. Hlawitschka, and J. Kennedy, editors, *EuroVis - Short Papers*. The Eurographics Association, 2014. 96

[125] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. Vega-lite: A grammar of interactive graphics. *IEEE Trans. Vis. Comput. Graphics*, 23(1):341–350, 2016. 22, 80

[126] Shrutika S. Sawant and Manoharan Prabukumar. A review on graph-based semi-supervised learning methods for hyperspectral image classification. *The Egyptian Journal of Remote Sensing and Space Science*, 2018. 95

[127] John Scott. Social Network Analysis. *Sociology*, 22(1):109–127, February 1988. 19, 29, 31, 39, 47

[128] Vanessa Serrano Molinero, Benjamin Bach, Catherine Plaisant, Nicole Dufournaud, and Jean-Daniel Fekete. Understanding the use of the vistorian: Complementing logs with context mini-questionnaires. In *Visualization for the Digital Humanities Workshop*, Phoenix, United States, October 2017. 37, 38, 39, 46, 55

[129] Rachel Shadoan and Chris Weaver. Visual Analysis of Higher-Order Conjunctive Relationships in Multidimensional Data Using a Hypergraph Query System. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2070–2079, December 2013. 62

[130] Termeh Shafie, David Schoch, Jimmy Mans, Corinne Hofman, and Ulrik Brandes. Hypergraph Representations: A Study of Carib Attacks on Colonial Forces, 1509-1700. *Journal of Historical Network Research*, pages 52–70 Pages, October 2017. 60

[131] Ben Shneiderman. Dynamic queries for visual information seeking. *IEEE Softw.*, 11(6):70–77, November 1994. 73

[132] Georg Simmel. *Soziologie: Untersuchungen über die Formen der Vergesellschaftung*. Duncker & Humblot, Berlin, 7. aufl edition, 2013. 32

[133] Marc A. Smith, Ben Shneiderman, Natasa Milic-Frayling, Eduarda Mendes Rodrigues, Vladimir Barash, Cody Dunne, Tony Capone, Adam Perer, and Eric Gleave. Analyzing (social media) networks with NodeXL. In John M. Carroll, editor, *Proceedings of the Fourth International Conference on Communities and Technologies, C&T 2009, University Park, PA, USA, June 25-27, 2009*, pages 255–264. ACM, 2009. 15, 38, 60, 69

[134] SNA — Tools for social network analysis.

[135] John Snow. On the Mode of Communication of Cholera. *Edinb Med J*, 1(7):668–670, January 1856. 20

[136] John T. Stasko, Carsten Görg, and Zhicheng Liu. Jigsaw: Supporting investigative analysis through interactive visualization. *Inf. Vis.*, 7(2):118–132, 2008. 39, 51, 55, 119

[137] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002. 97

[138] Shazia Tabassum, Fabiola S. F. Pereira, Sofia Fernandes, and João Gama. Social network analysis: An overview. *WIREs Data Mining and Knowledge Discovery*, 8(5):e1256, 2018. 11, 14, 31

[139] Natkamon Tovanich, Alexis Pister, Gaelle Richer, Paola Valdivia, Christophe Prieur, Jean-Daniel Fekete, and Petra Isenberg. VAST 2020 Contest Challenge: GraphMatchMaker: Visual Analytics for Graph Comparison and Matching. *IEEE Computer Graphics and Applications*, pages 1–1, 2021. 62, 63

[140] John W. Tukey. The Future of Data Analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962. 20

[141] John W. Tukey. *Exploratory Data Analysis*. Pearson, Reading, Mass, 1er édition edition, January 1977. 14, 23

[142] Paola Valdivia, Paolo Buono, Catherine Plaisant, Nicole Dufournaud, and Jean-Daniel Fekete. Analyzing Dynamic Hypergraphs with Parallel Aggregated Ordered Hypergraph Visualization. *IEEE Trans. Visual. Comput. Graphics*, 27(1):1–13, January 2021. 37, 53, 89, 120

[143] Guido van Rossum. Python tutorial. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995. 62

[144] Ingeborg van Vugt. Using multi-layered networks to disclose books in the republic of letters. *Journal of Historical Network Research*, 1(1):25–51, October 2017. 49

[145] Corinna Vehlow, Fabian Beck, and Daniel Weiskopf. The state of the art in visualizing group structures in graphs. In R. Borgo, F. Ganovelli, and I. Viola, editors, *Eurographics Conference on Visualization (EuroVis) - STARs*. The Eurographics Association, 2015. 96

[146] VisMaster: Visual analytics — Mastering the information age. 111

[147] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al. Constrained k-means clustering with background knowledge. In *Icml*, volume 1, pages 577–584, 2001. 95

[148] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, November 1994. 14, 31, 39

[149] Charles Wetherell. Historical Social Network Analysis. *Int Rev of Soc His*, 43(S6):125–144, December 1998. 11, 13, 19, 33, 41, 47, 48, 59

[150] Kai Xu, Alvitta Ottley, Conny Walchshofer, Marc Streit, Remco Chang, and John Wenskovitch. Survey on the Analysis of User Interactions and Visualization Provenance. *Computer Graphics Forum*, 39(3):757–783, June 2020. 63

[151] Michelle X. Zhou. "Big picture": Mixed-initiative visual analytics of big data. In *Proceedings of the 6th International Symposium on Visual Information Communication and Interaction*, VINCI '13, page 120, New York, NY, USA, 2013. Association for Computing Machinery. 96