université PARIS-SACLAY

Si la thèse a été préparée dans le cadre d'une **cotutelle internationale de thèse**, logo de l'université étrangère co-délivrant le doctorat **(sinon, à enlever)**

Titre de la thèse (sur plusieurs lignes si nécessaire)

Traduction du titre de la thèse (sur plusieurs lignes si nécessaire)

Thèse de doctorat de l'université Paris-Saclay et de l'université XXX (si cotutelle - sinon enlever cette seconde partie)

École doctorale n° d'accréditation, dénomination et sigle
Spécialité de doctorat : voir annexe
Graduate School : voir annexe, Référent : voir annexe

Thèse préparée dans la (ou les) unité(s) de recherche Nom(s) (voir annexe), sous la direction de Prénom NOM, titre du directeur ou de la directrice de thèse, la co-direction de Prénom NOM, titre du co-directeur ou de la co-directrice de thèse, le co-encadrement de Prénom NOM, titre, du co-encadrant ou de la co-encadrante ou la co-supervision de Prénom NOM, titre, du tuteur ou de la tutrice (en cas de partenariat industriel)

Thèse soutenue à Paris-Saclay, le JJ mois AAAA, par

Alexis PISTER

Composition du jury

Prénom Nom Titre, Affiliation	Président ou Présidente
Prénom Nom Titre, Affiliation	Rapporteur & Examinateur / trice
Prénom Nom Titre, Affiliation	Rapporteur & Examinateur / trice
Prénom Nom Titre, Affiliation	Examinateur ou Examinatrice
Prénom Nom Titre, Affiliation	Examinateur ou Examinatrice
Prénom Nom Titre, Affiliation	Directeur ou Directrice de thèse

THESE DE DOCTORAT

NNT : 2020UPASA001

université PARIS-SACLAY

ÉCOLE DOCTORALE
Physique et ingénierie : Electrons, Photons, Sciences du vivant (EOBE)

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

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

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

Title : titre (en anglais).....
Keywords : 3 à 6 mots clefs (version en anglais)

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

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

Table des matières

1 - Introduction

Social scientists such as historians and sociologists want to make sense of the structure and dynamics of the social relationships between people of a given place and time. Social Network Analysis (SNA) and its historical equivalent Historical Social Network Analysis (HSNA) is one of the main paradigms to achieve this task. It consists in modeling the social relationships between agents—such as persons or organizations— as a network and studying it to make sociological conclusions. Usually, agents are represented as nodes in the network, while the links model social relationships, such as friendships or family links. To construct such networks, social scientists try to exhaustively list all the persons in a restricted time and place with all their social ties and create a network from it. The resulting network is considered to be a good model of the social reality, thus allowing to study the structure and dynamics of the social fabric of a period, by studying the network in itself. In parallel, a lot of work has been done in network visualization and specifically Social Network Visualization (SNV) to make useful representations of social networks, and visual analytics tools allowing an effective exploration and analysis of this type of data. However, sociology and history data can be quite complex, and simple networks model are often a simplification of the real world social phenomena. Several network models ranging from simple to complex ones have been introduced, with associated visual representations and tools. Unfortunately there is no consensus on which model is best, specifically for networks constructed from historical sources, and the majority of research is still done using very simple models, with classical node-link representations. Furthermore, most widely used SNA tools such as Gephi or Pajek do not provide much guidance to the social scientists for their analysis, which often require good computer science and statistics skills. This thesis' aim is to tackle those two problems, by first defining a network model which models well most of the historical sources we encountered, and proposing visual representations to explore them. Then, we propose visual analytics tools and methods specifically designed for social scientists to explore their data, with the aim of proposing the right balance between algorithmic power, interpretation of the analysis and decision-making of the social scientist.

1.1 . Social History and Historical Social Network Analysis

History is the science of retrieving and characterizing facts about the past, in all their complexity. Traditional history methodology consists in finding and expliciting specific events—such as wars or diplomatic alliances—while eliciting their causes and consequences, and narrating the lives of historic figures, such as reigners and artists. But in the first half of the 20th century, in contrast of the general trend of event history a new methodology emerged, referred as social history. This branch

of history studies the socio-economic dynamics between the different groups of a society, instead of focusing on specific events and affairs of a state. More recently, with the development of network science and computer science, sociologists started to study social phenomena and relationships from a network perspective. A network is an abstraction based on graph theory concepts used to model phenomena based on relationships between entities, made of nodes and links. Sociologists started to use this concept to model social ties between agents of interest and study social phenomena through the description of the network structure, using the SNA methodology. It allowed them to leverage quantitative measures from the network to make sociological conclusions. This method has been used to study a variety of social constructions such as families, political institutions, schools, friendships, work environments, and sports clubs, with promising results. This network analysis approach grew in popularity in recent years, and has started to be used and formalized by historians, under the term of Historical Network Research (HNR). Similarly to sociologists, historians can build a network modeling the social relationships between actors of the past, restricted in a specific period and area they are studying. If sociologists can use surveys, experiments or nowadays the internet to extract social relationships and construct a social network, historians are restricted by the written sources they can find. Their main source of work to extract social relationships in a rigorous way are historical documents which correspond to traces of specific events linking people together. These can be marriage acts, birth certificates, census for family and close personal relationships, or migration acts and working contracts for other types of social ties. After having a selected corpus, they manually annotate each document to extract the persons mentioned in it along the relationships between them, to finally construct a network from this data. This is a long and tedious process which can result in small to large networks that they want to analyze to make conclusions on the social dynamics of a population of interest. For this complicated task, historians follow what is called a Social Network Analysis (SNA), or more precisely a HIstorical Social Network Analysis (HSNA) which consists in characterizing the structure of the network with measures such as the centrality or the density of some parts of the network to then make conclusions on how people were interacting in the period of interest. To help their analysis, and generate new hypotheses, they usually rely on Visual Analytics tools to represent and explore their network. The elaboration of visual tools to represent and explore social networks is called Social Network Visualization. Sociologists and historians started to use static representation of networks, using node-links diagrams to have a visual understanding of their data, and to report their findings in publications. With the development of visualization, more complex representations and visual analytics tools emerged, which allow more complex representation and exploration capabilities, with the help of interactions and navigations features. Social networks visual systems such as Gephi or Pajek are now widely used in HNR and SNA by social scientists. Representing their network data and being able to interact with

it allows them to rapidly have an overview of it, confirm hypotheses they have and arrange new ones by exploring the network.

However, most used social networks visual analysis tools still have several issues that we tackle in this thesis : the visual representations still widely used are pretty simple, and are often not a good fit to represent and explore complex multivariate historical dataset, and current visual analytics tools often do not provide enough power and guidance to the end users to manipulate their data, which can result in frustration.

1.2 . Network models and representations

Person-to-person simple node-link diagram is still the most widely used network representation is SNA, and most SVA tools only include this type of representation. This visualization shows the persons as nodes, and social ties as links and displays them in a way to minimize the number of crossings to increase the readability. However, historians very often have access to richer and more diverse information through the historical documents they study. The documents can refer to coexistent complex social relationships which link several people together with different roles. These cannot be modeled with simple person-to-person links, without losing some information on the social implication of these relationships. Moreover, documents often give access to other information related to the event they refer to, such as the time, the location or the roles of the different persons mentioned. For example, marriage acts often indicate the date and the place of the event, and mention persons under different roles : the spouses, the witness, the parents, the priest etc. Additional information related to persons can also be mentioned, such as their age, origin or profession. It is clear that simply using a person network model won't encapsulate the whole complexity of the data and will simplify the social relationship. This is a common issue in SNA and HSNA [REF Lemercier] and more complex network models are needed. However, complexity along with visual analysis tools to explore them.

1.3 . Usability Issues

One of the aims of Visual Analytics is to provide automatic or semi-automatic processing and analysis tools with data mining and machine learning algorithms, to help end users make sense of their data and find interesting patterns and relationships. However, current social network visual analytics systems are still very algorithm oriented, and do not provide many controls to historians and sociologists who usually feel off the analysis loop when the system provides automatic and algorithmic results. One of the reason is because automatic results can be hard to interpret, especially in a discipline such as History or Sociology, where users often have little knowledge on computer science. One example is the automatic detection

of community structures using network clustering algorithms. Social networks are known to have a community-like structure, meaning that the probability of a link existing between two random person nodes is not uniform, and that people tend to agglomerate in groups, who have more social ties between them than with other persons in the network. There are a lot of existing clustering algorithms which aim to automatically find these groups, by optimizing measures such as the modularity or using propagation models. However, clustering is an ill-defined problem, and several good partitions may coexist for the same network, and which can have several interpretations in a SNA. Most SNA/SVA tools such as Gephi or NodeXl provide several well-known clustering algorithms such as Girvan-Newman, Louvain or Clauset-Newman-Moore, but do not provide much guidance on how to use them and interpret their results. Social Scientists often try several ones in the list of algorithms proposed until finding a convenient result, in the eyes of the analysis they want to follow. This leads to a non-satisfactory analysis process, as historians are out of the loop and have few decisions on the results. This usability issue is the same for automatic processes with no universal ground truth.

1.4 . Contribution and research statement

This thesis is centered around two research questions : first, the proposition of an efficient network model to represent historical sources and follow HSNA, with associated visualizations to show and explore this type of network. Secondly, elaborating visual analytics tools specifically designed for this type of data, with the right balance between algorithmic power, simplicity and interpretability for the social scientists, who need to be in control of the analysis. We first [tell plan]

2 - Historical Network Analysis and Visualization

2.1 . Social Network Analsyis

2.1.1 . Sociometry to SNA

One of Sociology's main goal is to study social relationships between individuals and finding recurrent patterns and structures allowing to explain the behaviours of people and groups. Traditional methods try to explain social phenomena using classical social classifications such as the age, social status, profession and sex. For example, the social position of people living in a small city could be explained well by their age, demographics and social status which are traditional social categories. However, some criticism emerged that this type of division is often partially biased and come from predefined categories which are not always grounded in reality. Sociometry is considered as one of the basis of SNA and had the goal of redefining social categories through the lens of real social interactions and ties between persons, that sociologists wanted to observe in real conditions. It is in the 1930s that Moreno started to develop this new method by trying to depict real social interactions as a way to understand how groups and organization were functioning [**morenoFoundationsSociometryIntroduction1941**]. He elaborated sociograms as a way to visually show friendships between people with the help of circles representing persons and lines modeling friendships. This way, he could rapidly see the main actors and hubs of interaction inside the social network represented visually. Sociometry tremendously helped disseminate the metaphor of networks to model and understand social structures and phenomena. It was during the 1960s that sociologists took these concepts further and formalized SNA using graphs and mathematical methods, following the emergence of Graph Theory studies in the 1950 by Mathematician such as Erdyos . It did not take long until sociologists used these concepts to model social ties and relationships into graphs. Sociologist already have structural theories of social phenomena, and they rapidly saw the potential of graphs to model and analyze those in a mathematical way. Several sociologists started to codify those concepts to use them in a socioloogy setting such as Coleman (1964). They started to model social ties between agents as graphs $G = (V, E)$ with V a set a vertices representing agents such as persons and organizations, and $E \subseteq V^2$ a set of edges modeling the social ties between pairs of agents. Once they modeled social relationships as networks, a variety of mathematical methods coming from graph theory were at their disposal. Sociologists started to make links between these mathematical results and sociological facts. It was then possible to make sociological conclusions from the direct observations of social ties modeled as networks.

2.1.2 . Structuralism and Ego Studies

After SNA started to be formalized, lots of sociological studies have been done using those concepts. However, there was not yet strong protocols and methods to follow, and networks are an abstraction that can model different things in different ways. When looking retrospectielvy, we can see that two school of thougts emerged with different objective and methods : the structuralists and the school of Manchester.

The Structural Analysis of Social Network refer to the Structuralists in Sociology. They are interested in the proprieties and structure of the network, and make parallel between them and how persons were interacting in real life. They think the position of persons in the network and the relational patterns they are part of refclects well the social activities and behavior in real life. Accordingly, sociologists in this school usually study organisation and specific groups, and want to expian their behavior and interaction through the internal shape and structures of resulting networks. They thus try to contruct network which exhaustilvely model all the interactions between the actors constituing the groups.

In contrast, the school of Manchester try to explain specific persons behavior and social intertacions, through their direct interactions and without necessaritly studyin a global network structure. This school of thought is related to the concept of ego networks. Ego networks consists in all the direct relation of one node—in this case a personn—with the relation between persons of this small network. They usually want to model the different types of relationships of a person, like their family, work and friends ties and study them through time. They make a direct parallel between theses direct social ties and the status, condition and life of persons, and usually compare several ego networks to make conclustions about the correlations between the two.

These two ways of seeing SNA are often not exclusive and current studies usually involve concepts and methods from these two schools.

2.1.3 . Methods dans tools

Graph theorists and network scientists developed a myriad of measures and algorithms that sociologists appropriated themselves to describe and charactierize social phenomena. When constructing networks, the first thing sociologists did was often to identify the main actors of the network, and explain why these actors were the most central, for example by linking it to their profession or social status. Computing the degree—which is the number of connections—distribution is the main straightforward way of doing it, but other more complex measures like the centrality have been developed too. Lots of types of centrality have been proposed, based on different criteria, as there are several ways of defining the more <emph>important</emph> actors. Centrality can highlight actors with the highest number of connections while others highlight people bridging different groups with low interactions. More generally sociologists aimed at identifying recurring

patterns of sociability between actors. The concepts of dyads and triads counting which are simple structural elements give insight on that and reflects on Simmel formal sociology, where he already refered as dyads and triads as primal form of sociability. More recently, the concept of graphlet extended this concept to every pattern of N-entities. Graphlet analysis aims at enumeraing every small structure of N nodes composing a network, to understand how people interact at a low-level. Graphlets counting shows that graphlets are not found in an uniform distribution in social networks, thus revealing that these networks do not follow a random distribution. It is a fact well known by sociologists and more broadly every person working with real worl networks. More precisely, entities in real world networks tend to agglomerate into groups, where entities in the same groups interact more between them than with entities in other groups. In a sociology perspective, it means that people tend to interact and socialize in groups, and interact more rarely with other people. These groups are often refered as <emph>communities</emph>, and a lot of algorithns have been proposed to find these automatically.

2.2 . Historical Network Research

2.2.1 . Social History

Historians try to understand an epoch using textual sources from the past, and trying to extract useful information from them. Social history, which is a branch of history, focus on understanding how societies were organised and how people were living together at a particular time and place. Charles Tilly argued that the task of social history lays in "(i) documenting large structural changes, (2) reconstructing the experiences of ordinary people in the course of those changes, and (3) connecting the two". For the latter, historians can leverage personal written sources—such as letters, journals, books, and newspapers—to have the internal point of view of persons living in this society and descriptions of lives of precise individuals. For the former, historians usually need to study more structured documents which contain information which can be extracted in a predefined and exhaustive way. These documents can for example be census, migration acts or marriage acts. By studying theses documents and by systematically extracting the information of these documents, historians can make global and quantitative conclusions on certain social and behavioural aspects of societies of interest. For example .. [EXAMPLE CHANGEMENT METIERS XXth century]

2.2.2 . Historical Social Network Analsyis

History started to adopt some of the methods and vocabulary of Network research in the 1980s, several years after other fields such as Sociology or Anthropology (TO CHECK). Before that, historians were already describing relational atructures when studying families and organization. It was often a part of discussion and a conclusion of several studies. Network reseach was a way to put

these relational structures as an object of study in iteself, and allowed to study them in a more systematic and quantitative way. Instead of only looking at classes and groups, historians thus started to look at relational links between individuals, such as family, friendships or business ties. They already had techniques and tools to annotate and extract quantitative information from textual sources that they adapted to extract and study social ties. We therefore saw the emergence of HNR studies, where historians followed HSNA studies on networks constructed from the mention of social ties of their textual sources. It allowed them to make observations on previous objects of study like families or organization that it was not possible to see without taking into account the relational aspects of these phenomena. However, constructing a network from historical sources, which can differ in their structure is not a trivial task. The most straightforward approach, based on the most well known social network analysis, consists in constructing social network based on simple graph $G = (V, E)$ with V a set a vertices representing the persons of interest, and $E \subseteq V^2$ a set of edges modeling the social ties between pairs of persons. This allows to have a simple network to visualize and analyze, but does not always reflect the social complexity of the real relationships. More complex networks models have been proposed in SNA to be able to model more complex social relationships.

2.2.3 . Network Modeling

The (H)SNA network models have evolved over time to better take into account concrete properties of social networks, such as types of actors using labeled networks, the importance of actors or relations with weighted networks, mixed relationships with multiplex networks, dynamics of relations with dynamic networks. Bipartite networks have been proposed to model relations between two types of entities, such as organization and employees where the relations link employees to organizations but not employees to employees or organizations to organizations. Many social situations or documents can be modeled in these terms (affiliation lists or co-authoring). Multivariate networks, i.e., graphs, where vertices and edges can be assigned multiple "properties" or "attributes", are less used in SNA. These attributes are often considered secondary, the emphasis of SNA being on the topology, its features, measures, and evolution.

Historians, demographers, sociologists, and anthropologists have been designing specific data models for their social networks, based on genealogy or more generally kinship [**hamberger:halshs-00658667**]. For genealogy, the standard GEDCOM [**gedcom**] format models a genealogical graph as a bipartite graph with two types of vertices : individuals and families. This format also integrates an "event" object but it is diversely adapted in genealogical tools. The Puck software has extended its original genealogical graph with the concept of "relational nodes" to adapt the data model to more family structures and to integrate other social relationships for anthropology and historical studies [**hamberger_scanning_2014**].

2.3 . Social Network Visualization

2.3.1 . Visualization

Visualization consists in graphically displaying data in the purpose of enhancing human cognition capabilities to understand and communicate ideas and phenomena. History is filled with classical examples of visual data display which helped understand real phenomena, such as Minard's map of Napoleon march in Russia, or the cholera crisis. Visualization then developed mainly from the 1960s as a research field with the rise of computer science and hardware capabilities. As the amount of data stored increase exponentially, descriptive statistics were not enough to understand the underlying structure of the amount and diversity of produced data. Visualization, leveraging the human visual system, allows to rapidly see the structure of a dataset and detect interesting and unexpected patterns VERY often unseen with classical statistical methods. One famous illustration of this is Ascombe quartet, four datasets with the same statistical values but with very different structures, that plotting the data highlight. Lots of visualization techniques emerged to make sense of the diversity of data produced, such as relational, temporal, spatial or network data. More precise taxonomy then emerged : **Scientific visualization** focus on visualizing continuous real data such as weather, spatial, and physics data, sometimes produced with simulations whereas **Information Visualization** is centered around visualization (multidimensional) discrete data points, often in an abstract way. Semiology of graphics Grammar of graphics

2.3.2 . Social Network Visualization

Sociologists rapidly saw the potential of graphically showing relationships between individuals, to better comprehend the underlying social structure and communicate their findings. Moreno elaborated sociograms to visually show friendships among schoolers with circles and lines to respectively show children and friendships ties. This type of representation—commonly called node-link—is the most widely used in social sciences, as it is rapidly understandable and effective for small to medium-sized networks. Finding an optimal placement for the nodes is however not that simple as several metrics can be optimized depending on the desired drawing, such as number of edge crossings, the variance of edge length, orthogonality of edges etc. The number of edge crossings is often considered as the most important measure, but finding a drawing with the optimal number of crossing is a NP-Hard problem, meaning that heuristics are needed for most real world use cases. Lots of algorithms have been designed such as force-directed ones, modeling the nodes as particles which repulse each other and are attracted together when connected with a link. Other visual techniques have been proposed to represent network such as matrices and arcs, but are less used in social sciences. Still, Matrices have been shown to be better than node-link diagram for a lot of tasks such as finding cluster related patterns, especially for medium to large networks. As social scientists started to use more complex network models such as bipartite or temporal net-

works, more sophisticated representation are needed. The visualization community proposed new visualization systems for specific network types such as PAOHVis for temporal hypergraphs, NodeTrix for clustered networks or Juniper for Multivariate networks. However, these new networks representations take time to be adopted by social scientists, and rarely use those.

Moreno 1930 Node-Link NP complet Heuristiques Mesures (croisements/taille des liens etc) Autres techniques Arcs Matrices Autres graphes Temporel Hypergraphs

2.3.3 . Social Metwork Visual Analytics

Visualization has mostly been used for confirmatory and communication purposes from its beginning. Social scientists often had hypothesis that they could rapidly verify by plotting the data. The same plots were often used for communication purposes, for example in a scientific paper or presentation. However, visualization can also be used for exploratory aims, to gain general insight on the data and potentially generate new hypothesis. This process has been characterized by Tukey in 1960 as **exploratoy data analysis**. Exploration is mostly possible thanks to interaction mechanisms, which allows to change the point of focus in the data to highlight interesting patterns, with the help of mechanisms like filtering, querying, sorting etc. As the average size of datasets keeps growing, exploratory tools are often needed. However, social scientists often have hypothesis they want to verify, even before plotting their data. They also sometimes want to gain insight with the help of statistical and machine learning methods, that visualization only can not provide. More recent visual exploration interface incorporate analytical tools with the visualization, letting users apply statistical or machine learning algorithms directly in the exploratory loop. This coupling of visualization and analytical reasoning has been defined as Visual Analytics (VA) and is still undergoing lots of research. Social scientists now frequently use VA systems to explore their data and apply statistical and machine learning algorithms to verify and create hypothesis. Unfortunately, social scientists are often not trained in computer science and mathematical methods, and a lot of them have been frustrated by VA tools by how it was guiding their analysis in predefined ways. For example, lots of social network VA interfaces propose clustering features, allowing users to find interesting groups with the help of automatic algorithms. However, social scientists often do not understand how the algorithms work and are not always satisfied with the results, as they can have knowledge from other sources not modeled inside the network. Cleaning and importing the data is also complicated, as the modeling process is not straightforward and social scientists often encounter errors in the data once they visualize it, that they would like to correct. Modern social network VA tools should support those tasks.

3 - HSNA Process and Network Modeling

3.1 . Introduction

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, social history needs to generate many networks from the same documents/sources to visualize and analyze them. In this article, after describing the current Historical Social Network Analysis [**wetherell_historical_1998**] (HSNA) workflow, we explain why and how effective tools for supporting this process should model social networks in multiple steps to support three essential principles : *traceability*, connection to *reality*, and *simplicity*.

Social historians' goal is to characterize socio-economic phenomena and their dynamics in a restricted period and place of interest, to see how individual people of that time lived through those changes. For this, they rely on historical documents such as conversational letters, censuses, and marriage acts. They usually extract qualitative and quantitative information from an identified corpus of documents, to then make conclusions on interesting socio-economic topics such as migrations, business dynamics, education, and kinship. For doing this, historians can apply Social Network Analysis (SNA), a method—sometimes referred to as a paradigm—which consists in modeling the social relationships between a set of entities—usually individuals—into a network. Much work has been done to adapt SNA to the context of historical document exploitation, and although several approaches coexist they can be brought together under the banner of Historical Social Network Analysis [**wetherell_historical_1998**] (HSNA) or Historical Network Research [**kerschbaumerPowerNetworksProspects2015**] (HNR). When following an HSNA, historians collect documents, annotate them, and construct a network from the annotations that they finally analyze and visualize to validate or find new hypotheses. Unfortunately, the process is often linear and it is common that, when visualizing their network, historians spot errors and inconsistencies in the annotations that they could have fixed if the process was iterative.

Moreover, historical documents are often complex and the annotation and modeling process can be done in many different ways. Several network models have been proposed ranging from simple and specific ones like co-occurrence networks to more general and complex ones such as multilayer networks and knowledge graphs. Simple models allow answering specific questions and are easy to manipulate but are often too simplistic and may distort the information contained in the documents. Moreover, they often break the traceability from the analysis to the original documents, making the process of cleaning the annotations complicated. Indeed, errors and mismatches often occur in the annotation process, for example,

due to entity disambiguation problems. On the contrary, too complex models are complicated to visualize and analyze, and historians do not always have the tools to create them properly. This paper proposes to model historical datasets as bipartite multivariate dynamic networks, where both persons and documents are modeled as nodes with attributes. While this model is simple enough for creation and inspection, it allows tracing back the entities of the network to the original sources for a continuous annotation process and still accurately models the social relationships mentioned in the documents. Historians can therefore use this model to simultaneously find errors and inconsistencies in their annotation process—allowing them easier back and forth between the annotation and analysis steps—while starting a first analysis and exploration of the data to answer their sociological questions.

3.2 . Related Work

We summarize here work bridging social history to network analysis and visualization.

3.2.1 . Quantitative History

Traditionally, historians try to tell a story about protagonists and socio-economic facts in a given society by reading, understanding and linking together historical sources. This narrative approach to history has been criticized for its lack of traceability and the open interpretation of historical documents, which can introduce bias from the author. To solve this problem, the "Annales school" (Ecole des Annales) proposed to characterize past social phenomena through the exhaustive and systematic analysis of historical documents [**prostDouzeLeconsHistoire2014**]. Quantitative approaches then emerged in the 1960s with the appropriation of statistical and computer science methods to analyze data extracted from historical documents. This is the case of Historical Demography which works on nominative data to produce quantitative results on fertility and mortality [**henryRegistresParoissiauxHistoire1956**]. Unfortunately, these approaches have been criticized for their simplifications and for consuming considerable time while often providing simple results [**karila-cohenNouvellesCuisinesHistoir**]. Approaches using digital methods and tools are nonetheless more and more popular, sometimes referred to under the umbrella term Digital Humanities. If their adoption remains slow and sometimes criticized among historians, they still provide tools to store, explore, and analyze historical documents systematically if used appropriately (i.e. not trying to bias the analyzes). It can also provide infrastructure and tools to study large historical databases, as with the Venice Time Machine project [**kaplanVeniceTimeMachine2015**] 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.

3.2.2 . Social Network Analysis and History

In Sociology, networks have been a common metaphor to talk about social relationships [**freeman_development_2004**] which can be easily thought of as invisible bonds linking individuals, forming a global structure of connections similar to a mesh or a web. After extensive work in graph theory and network modeling developed in the 1950s, anthropologists and sociologists started to borrow those concepts from maths and use them to model social relations—such as family, friendships, or business ties—with networks [**coleman_introduction_1964**, **freeman_development_2004**]. SNA revolutionized classical Sociology by trying to explain social phenomena through the lens of real interactions modeled as networks, while classical methods were revolving around predefined social groups such as age and gender.

History started to use those concepts and methods in the 1980s [**wetherell_historical_1998**] in order to criticize quantitative history concepts and results, and to develop historical approaches—like *Microstoria* [**ginzburgMicrohistoire1981**]—that focus on the study of individuals and groups through the lens of their interactions and relationships directly extracted from historical documents. Since then, HSNA has been applied by sociologists and historians to study multiple kinds of relationships, like kinship and political mobilization [**lippKinshipNetworksLocal2005**], administrative and economic patronage [**moutoukiasReseauxPersonnelsAutorite1992**], etc. If these approaches fall under similar critics of quantitative history [**lemercier12FormalNetwork2015**] lots of historians are using and continuously improving this method which can be very effective to study relational historical phenomena [**kerschbaumerPowerNetworksProspects2015**]. Moreover, historians rarely rely on a single approach when studying an era or phenomenon, they mix methods and tools from several domains of social and formal sciences with their own practices [**padgettRobustActionRise1993**, **petzCombiningNetworkResearch2**

3.2.3 . Network Modeling

Social scientists started to model social relationships using simple graphs $G = (V, E)$ with V a set of vertices representing actors—very often persons—and $E \subseteq V^2$ a set of edges modeling a social tie between pairs of actors [**freeman_development_2004**].

The (H)SNA network models have evolved over time to better take into account concrete properties of social networks, such as types of actors using labeled networks, the importance of actors or relations with weighted networks, mixed relationships with multiplex networks, dynamics of relations with dynamic networks. Bipartite networks have been proposed to model relations between two types of entities, such as organization and employees where the relations link employees to organizations but not employees to employees or organizations to organizations. Many social situations or documents can be modeled in these terms (affiliation lists or co-authoring). Multivariate networks, i.e., graphs, where vertices and edges can be assigned multiple "properties" or "attributes", are less used in SNA. These attributes are often considered secondary, the emphasis of SNA being on the topology, its features, measures, and evolution.

Historians, demographers, sociologists, and anthropologists have been desi-

gning specific data models for their social networks, based on genealogy or more generally kinship [**hamberger:halshs-00658667**]. For genealogy, the standard GEDCOM [**gedcom**] format models a genealogical graph as a bipartite graph with two types of vertices : individuals and families. This format also integrates an "event" object but it is diversely adapted in genealogical tools. The Puck software has extended its original genealogical graph with the concept of "relational nodes" to adapt the data model to more family structures and to integrate other social relationships for anthropology and historical studies [**hamberger_scanning_2014**].

3.2.4 . Social Network Visualization

Social scientists such as sociologists and historians always used visual representations for social networks [**cristofoliPrincipesUsagesDessins**], mainly for communication purposes and sometimes for exploration [**brandesExploratoryNetworkVisualization**]. Moreno elaborated on sociograms in the 1930s to visualize friendships using circles and lines to represent persons and ties, in a node-link fashion [**moreno_foundations_1941**]. Node-link diagrams are still the most widely used technique in SNA and HSNA by far to represent networks, despite scalability issues. The most used social network visual analytics software such as Gephi [**Gephi**] and Pajek [**batagelj_pajek_nodate**] are based on this type of representation and allow a fully integrated exploration and analysis with the help of various algorithms. The visualization community also proposed other representations to visualize networks such as matrices [**behrischMatrixReorderingMethods2016**] and arcs [**dangTimeArcsVisualizingFluctuations2016**], and to explore other network types such as dynamic hypergraphs with PAOHVis [**valdivia_analyzing_2021**], clustered graphs with NodeTrix [**henry2007nodetrix**], geolocated social networks with the Vistorian [**vistorian_mini_questionnaires**], and multivariate networks with Juniper [**nobreJuniperTreeTable2019**]. Jigsaw [**Stasko**] is designed to analyse a collection of documents. It encompasses the gathering of documents, named entity recognition, with analysis and visualization methods ; the last three steps of our workflow. These documents and entities form a multi-partite network with a few attributes (e.g., document title, date) but without roles. Except for Jigsaw [**Stasko**], most of the tools proposed are solely focused on the exploration and analysis of the final network and do not take into account the context of the whole HSNA process which led to the network creation.

3.3 . Historical Social Network Analysis Workflow

The essence of the Historical discipline is based on a critical approach of sources and involves considering peers' work. Traditional approaches to history often focus on the construction of a narrative, without necessarily adopting a systematic and problematized approach to the exploitation of original sources. Social history and the "Annales School" brought answers to this problem by proposing to rigorously extract information from historical documents and make conclusions from them. Similarly, Glaser and Strauss developed the "Grounded Theory"

[**glaserDiscoveryGroundedTheory2010**] as a methodology for the humanities to build hypotheses and theories by solely studying and categorizing real-world observations, without starting from prior knowledge and predefined categories. Later on in the 1960s, quantitative methods started to be used in history, providing statistical and later computer-supported tools to aid historians in grounding their analysis in mathematical models and methods. Unfortunately, the lack of methodology and understanding between the two worlds led to many criticisms by historians pointing to using wrong metrics, simplifying categories, and disconnections between the original documents and analysis [**karila-cohenNouvellesCuisinesHistoire2018**, **lemercierBackSourcesPracticing2021**]. HSNA, which can be seen as a subcomponent of quantitative history has been criticized for similar reasons [**lemercier12FormalNetwork201**
Still, the usage of networks for historical analysis provided some interesting results and classical works [**padgettRobustActionRise1993**], meaning that a clearer and more rigorous methodology with simple tools grounded in the historian workflow and sources could improve the methods of the field. Karila-Cohen and al. provide advice on how to use quantitative methods in history [**karila-cohenNouvellesCuisinesHistoire2018**] while Dufournaud describes her workflow when studying the socio-economic status of women in France in the 16th and 17th centuries, which she splits into three steps : *data collection*, *data processing*, and *data analysis* [**dufournaudCommentRendreVisible2018**].
From the literature and discussions with historians' collaborators, we propose an HSNA workflow divided into 5 steps : *textual sources acquisition*, *digitization*, *annotation*, *network creation*, and finally *visualization and analysis*. The workflow is presented in **??**along with recurrent pitfalls.

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)**.

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.

Annotation Annotation is the process of finding and extracting useful information from the documents concerning the persons, their social ties, and any useful information for the historian. This extra information can concern the persons (their age, profession, sex, ethnicity, etc.) and their social relationships (type, date, place). It encompasses named-entity recognition (NER) as well as their resolution. Historians also sometimes annotate information on other entities mentioned in the documents, such as art objects or administrative entities. Usually, historians have a first idea of what they want to annotate in the data as they already explored the documents beforehand and have knowledge of their subject of study, with hypotheses they want to explore. It is however common they can change their mind through the annotation process, by reflecting on what they found in the documents. Unfortunately, this can produce **missing annotations (P3)** and **inconsistent annotations (P4)** at the end of the process if annotators are not careful. This task can also be challenging and the choice of annotations has an impact on the final network. Historians also face ambiguity in the process, as several different 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 [**andrei2011porgy**]. However, errors in the disambiguation process can lead to important distortions in the final network structure and properties [**diesnerImpactEntityDisambiguation2015**], 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) [**TEI**] 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 [**dufournaud:hal-00876586**, **vistorian_mini_questionnaires**]. 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.

Network Creation Historians construct a network from the annotations of the documents. Usually, all persons mentioned are annotated and will be transformed into network nodes (vertices). Additional information such as their age, profession, and gender can be stored as node attributes. How the network's links are created is not as trivial and can vary from project to project [**alkadi2022**]. The most straightforward approach is to create a link between every pair of persons mentioned in one document, thus forming a clique motif. This is a simplistic heuristic as social relationships can be quite complex, involving more than two persons who can have different roles in the relationship. The choice of the network model has a major impact on the future analysis and **may add bias if chosen loosely (P6).** More complex models have been proposed in the literature such as weighted, dynamic, bipartite, and layered networks.

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 [**freeman_development_2004**]. Usually, historians start to represent their network to visually confirm information they know, then to gain new insight with exploration. Representations need to be chosen wisely given the network as **some insight may be seen only with some specific visualization technique (P7)**. To test or create a new hypothesis, historians usually rely on algorithms and network measures. **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).

3.4 . Network modeling and analysis

Historians usually 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 [**dufournaudCommentRendreVisible2018**]. 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 scien-

tists 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. Moreover, network models satisfying *traceability*, *reality* and *simplicity* properties would mitigate those problems by allowing to navigate more easily between the network and the documents while still modeling well the social relationships mentioned in the sources and being easy enough to visualize and manipulate to detect potential errors and inconsistencies.

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 [wetherell_historical_1998] :** 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 [sairioMethodologicalPracticalAspects2009] :** Only the persons are represented as nodes, and two persons are connected with a link when they are mentioned in the same document (or section). This is a simple model and one of the first to have been used in SNA and HSNA. The major drawback of this model is that it does not take into account the diversity of social relationships, as every link is identical. It can work well when only one type of social relationship is studied like a friendship network [**moreno_foundations_1941**]. However, historical documents rarely mention only one type of relationship and this model is thereby very limiting for HSNA.

— **Multiplex Unipartite Networks [eriksonMalfeasanceFoundationsGlobal2006b] :** 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 e.g., as parent, friends, and business relationships. 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 [hamberger_scanning_2014] :** Nodes can have two types : persons and documents in this network model. A link refers to a mention of a person in a document and can thus only occur between persons and documents nodes. Usually, links are not typed and only encode mentions. More recent analysis in HSNA encode the *roles* of the persons in the documents as link types [**Cristofoli2018**]. This network model is more aligned with the original sources and allows following an analysis through the original documents themselves and not through concepts. For example, the GEDCOM format introduces the concept of "family" that ties together a husband, spouse, and children with different link types. However, the concept

of family can have different meaning across time and cultures, meaning that GEDCOM adds a conceptual layer instead of grounding the network to concrete traceable documents and events (e.g., no marriage but birth certificates).

— **Multilayer Networks [multilayer] :** 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 [**CRNOVRSANIN201456**] and historians [**vanVugt_2017**], 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)[kgraphs] :** 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 can rank the models given two axes : simplicity/complexity and specificity/expressiveness. Currently, historians mostly construct unipartite networks (simple, co-occurrence, and weighted) which are simple and allow them to answer specific questions. However, those models do not capture all the complexity of the documents and social scientists may miss important patterns. For example, modeling only co-occurrences of persons in documents remove the variety of social relationships these mentions can refer to. Several interpretations may coexist to explain why someone is central in the resulting network, which may be impossible to validate without encoding more information—such as the types of relationships—in the model. Depending on the schema of the annotations, it may be impossible to create more complicated networks at this step without redoing the annotation process which is costly in time and resources. On the contrary, too complicated models such as KG are difficult to create from the sources and are hard to visualize and analyze. Therefore, we argue that historians should aim to model a network that is simple enough to manipulate, can be traced back to the original sources, and model well the social reality of the documents—i.e. having those three properties : *simplicity*, *traceability*, and *reality*.

3.4.2 . Bipartite Multivariate Dynamic Social Network

We argue that historical documents are well modeled by bipartite multivariate dynamic networks, which have the following properties :

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

Links and Roles : A link models the mention of a person in a document. **Each link has a type corresponding to the roles of the persons 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 [**Stasko**] does not consider the roles.

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

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

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

One of the main benefits of this model is that the document nodes represent both the physical documents as well as 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.). This model is *simple* enough to manipulate and visualize for historians and allows tracing back every entity of the network to the documents according to the *traceability* principle. Still, the network preserves the *reality* of the social relationships mentioned in the sources. More precisely, Cristofoli demonstrates that bipartite networks ensure no distortion or ambiguity, unlike projected networks when modeling textual sources [**cristofoli_aux_2008**]. Furthermore, when attributes are encoded, projected networks provoke a duplication of information related to the events. For example, the date of a marriage can be encoded in the document/event node with a bipartite network while the same information would have to be stored in several links when using a projection.

3.5 . Applications

Several tools have been designed for visualizing dynamic bipartite networks that can also be considered dynamic hypergraphs [**valdivia_analyzing_2021**, **penaarayaHyperStorylinesInteractivelyUntangling2022**], but few incorpo-

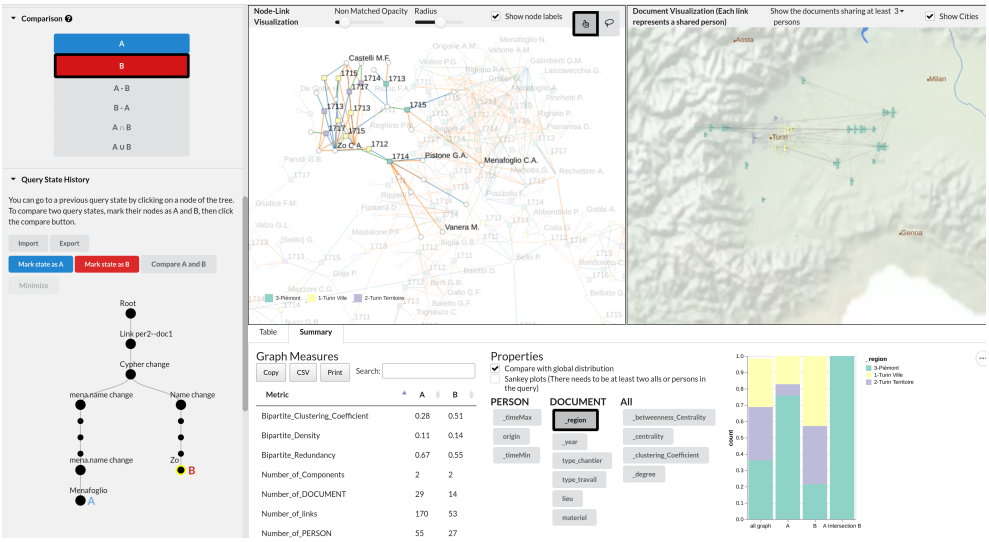


Figure 3.1 – ComBiNet interface exploring construction contracts in Piemont during the 18th century [**Cristofoli2018**]. The left menu lets users filter the data and compare groups. The center view shows the bipartite network with a node-link diagram. The right view shows a map with the geolocated construction contracts. The bottom view gives measures and attribute distributions related to the network and the current filters and comparisons. The user currently compares the *Menafoglio* and *Zo* families in terms of their construction types and close relationships.

rate attributes and complex interactions. We designed ComBiNet [**pister2022visual**] to explore and navigate through historical documents modeled as bipartite multivariate dynamic networksand to help social scientists answer their questions with the help of visual queries and interactive comparisons of query results. Figure 3.1 shows the interface to compare two meaningful groups of construction documents in Piemont during the 18th-century [**Cristofoli2018**]. In this example, we see that the *Zo* family has more construction contracts in *Turin* than the *Menafoglio* family. Exploring historic datasets modeled as bipartite multivariate dynamic networksallows answering complicated questions both related to the events (here the constructions) and the persons while being able to trace back to the original documents directly in the interface for cleaning or debugging purposes.

3.6 . Discussion

Most tools for social network visualization focus solely on the visualization and analysis steps, without considering the whole historical data analysis process, preventing researchers from going back to the original source, and supporting the social analyst in the annotation and modeling steps. We think visual analytics tools

helping social scientists annotate and model their data with *reality*, *traceability*, and *simplicity* principles in mind are essential to conducting socio-historical inquiries with limited friction, realistic training, and scientific transparency. Concerning the network modeling step, the bipartite multivariate dynamic networksmodel well the majority of structured historical documents such as marriage acts, birth certificates, and business contracts as these documents refer to specific events (birth, marriage, transaction, etc). The document nodes, therefore, represent both the textual documents and the specific events. This dual representation works well for semi structured document but could be more limiting for other more literary documents. Moreover, structured documents can also provide information about other relationships not directly linked to the main event. For example, marriage acts sometimes refer to the place and date of birth of the spouses with the names of the parents. 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), turning the network into a model of the documents and events and not of events only anymore.

3.7 . Conclusion

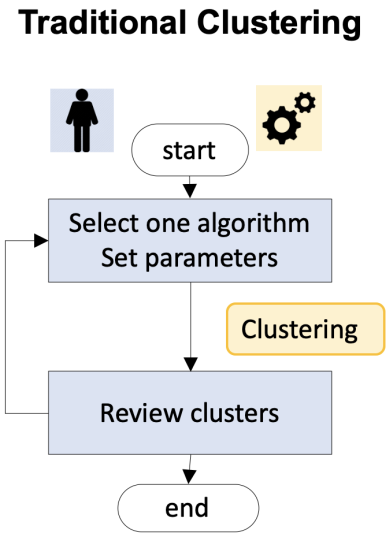
HSNA is a complex process that starts by collecting historical documents and ends with elaborating high-level sociological conclusions. Historians support their conclusions by modeling individuals' social relationships extracted from the documents and analyzing the resulting networks. We tried to shed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

4 - PK-Clustering

4.1 . 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 [**Rstat**], Gephi [**gephi**], Python with NetworkX [**networkx**], or Pajek [**pajek**]. To compute clusters, they follow a strained process : they select and run algorithms provided in the tool and then try to make sense of the results (see section 4.1). When they are not satisfied or unsure, they iteratively tweak the parameters of the algorithms at hand, run them again and hope to get results more aligned with their prior knowledge. This analysis process is un-



satisfactory for three main reasons : r0.5
Traditional Clustering. The output is a clustering, usually from a randomly chosen algorithm.

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.

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 [**kleinberg2003impossibility**]. 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 lets 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 4.1) :

1. *Specify Prior Knowledge (PK)*. Users introduce their prior knowledge of the domain by defining partial clusters. The tool then runs all available clustering

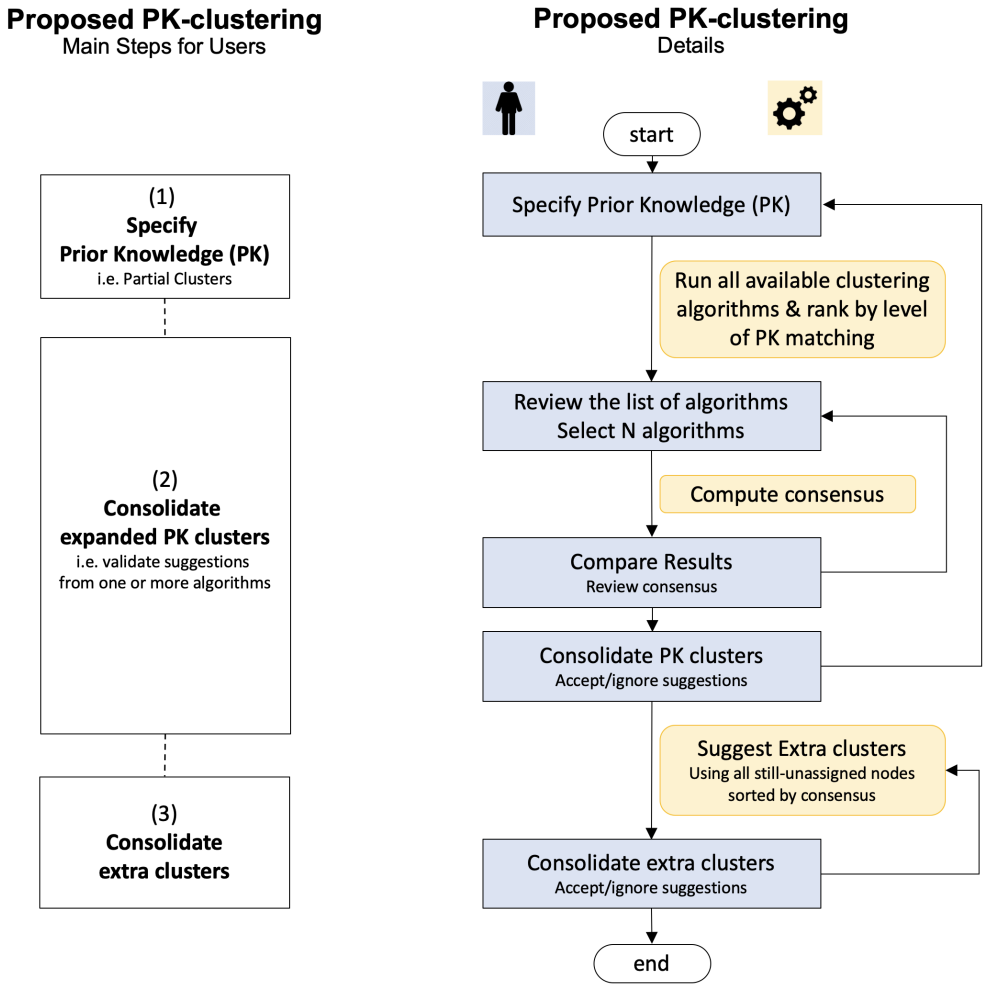


Figure 4.1 – PK-clustering. The output is a clustering supported by algorithms and validated (fully or partially) according to the user's Prior Knowledge.

algorithms.

2. *Consolidate expanded PK clusters*. Users review the list of algorithms, ranked according to how well they match the prior knowledge. They compare results and consensus, then accept or ignore suggestions to expand the prior knowledge clusters
3. *Consolidate extra clusters*. The tool suggests extra clusters on unassigned nodes. The user reviews consensus on each proposed cluster, then accepts or rejects suggestions.

The output of the process is, using a direct quote from a social scientist providing feedback on the prototype : "a clustering that is supported by algorithms and validated, fully or partially, by social scientists according to their prior knowledge".

According to the need to combine data mining with visualizations [**Ben02DiscoveryTools**] and inspired by the idea of letting the user collaborate with the machine to reach specific goals [**Horvitz99**], the proposed approach follows a user-initiated mixed-initiative [**Horvitz99**] visual analytics process.

In our case, users focus on the results that expand on their prior knowledge, filter-out the most implausible results, but can readjust when they realize that several algorithms are consensual despite not matching the prior knowledge (hinting at other possible meaningful structures). Our mixed-initiative approach allows social scientists to seed the clustering process with a small set of well-known entities that will be quickly and robustly expanded into meaningful clusters (details in subsection 4.3.1).

Contrary to a current trend[**molnar2019**], 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.

Currently, two types of social network analysis systems exist :

— interactive systems based on visualization and direct manipulation where researchers can tweak the layout and visual attributes to highlight communities ;

— data analysis systems where algorithms are used to compute "clusters" with various levels of control through "parameters"

These two worlds are still quite separated and they offer a trade-off between four aspects : control, speed, understandability and reproducibility. Manual systems offer a complete control to the analyst and the outcome of the work is understood, at least by the persons who created the communities, but this work is very slow, tedious, and probably not reproducible (another set of researchers will produce different communities). Automatic systems are fast and reproducible, but provide very little and indirect control (parameter tweaking) and their results is often difficult or impossible to understand.

Our main question is : Can we bridge the two worlds : providing more control to the users, higher speed, and yet keep the results understandable and the process reproducible ?

4.2 . 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.

4.2.1 . Graph Clustering

One of the main properties of social networks is their community structure [**Girvan7821**] 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, algorithms which use only the structure of the network to find clusters. [**FORTUNATO201075**] 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 [**alzahrani2016community**].

Moreover, recent new approaches try to use the attributes of the nodes [**yang2013community**]

and the dynamic aspect of the networks [**rossetiSurvey**] to find more relevant communities. Some toolkits offer a large number of algorithms ; for example, the Community Discovery Library (CDLIB) [**cdlib**] implements more than 30 clustering methods with variations inspired by 67 references.

4.2.2 . 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 [**basu08**, **wagstaff2001constrained**]. 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 [**basu02**], Fuzzy-CMeans [**bensaid96**], hierarchical clustering [**bohm08**], Density-Based Clustering [**lelis09**], and graph-based clustering [**wagstaff2001constrained**]. Shang et al. [**shang2017efficiently**] 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 [**SAWANT2018**] 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.

4.2.3 . Mixed-Initiative Systems and Interactive Clustering

Introduced by Horviz [**Horvitz99**], mixed-initiative systems are "interfaces that enable users and intelligent agents to collaborate efficiently". Several Visual Analytics systems are based on mixed-initiative interactions, [**makonin16**, **cook15**, **zhou13**, **wall18**], in particular the interactive clustering systems.

PK-Clustering is an interactive clustering system. A review by Bae et al [**baeetal20**] 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 ([**cavallo2018clustrophile**, **l2015xclusim**]) 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.

4.2.4 . 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. [**EVstar.groupstructures15**]. 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 [**Jianu14**]. 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 [**Saket14**], using new tasks assessing group-level understanding.

Holten [**Holten2006HierEdgeBundles**] 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 [**proulx06sandbox**]. A lot of work has also been done on the visualization of categorical variable in tabular data [**kosara2006parallel**, **gratzl2014domino**], which is similar to the notion of groups in networks.

4.2.5 . Ensemble Clustering

In the context of machine learning, an ensemble can be defined as "a system that is constructed with a set of individual models working in parallel whose outputs are combined with a decision fusion strategy to produce a single answer for a given problem" [**wang08**]. Several strategies exist for combining multiple partitions of items in a clustering setting [**strehl2002cluster**]. Concerning visualization research, Kumpf et al. [**kumpf18**] consider ensemble visualization as a sub-field of uncertainty visualization, for which some surveys exist [**Bonneau2014**, **maceachren05**]. 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.

4.2.6 . 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.

4.3 . PK-clustering

We present a new approach, inspired by the three types of clustering methods described in subsection 4.2.6 : 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 tool [**paohvis**] 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.

4.3.1 . 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 4.1. 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, 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 (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.

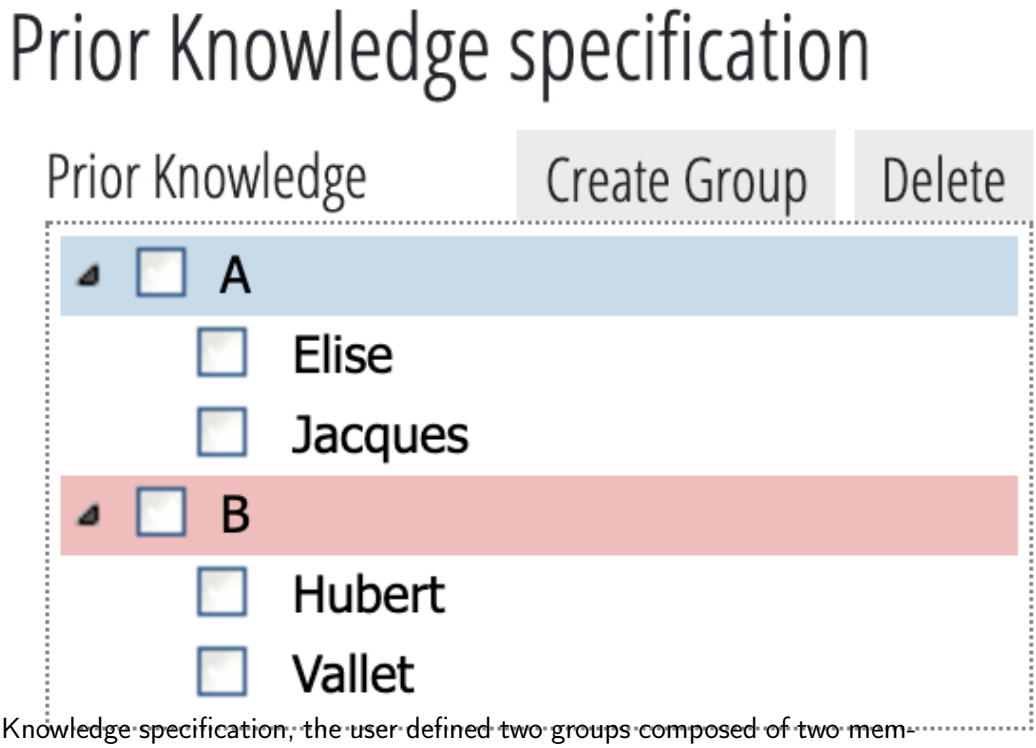
(2) Consolidate expanded PK clusters.

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 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.



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

4.3.2 . 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 item 4.3.1, each of the two prior knowledge groups contains two nodes, and it specifies that the user is expecting to see at least two clusters, with the first two people in a blue cluster A, and the other two in a red cluster B. This representation expresses *must-link* and *cannot-link* constraints described in subsection 4.2.2 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.

4.3.3 . 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 (male / female) it finds matching attributes and merges them if necessary. For numerical attributes (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 [**baroni17**] and try to optimize a topological measure such as *modularity* [**brandes08**]. 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 [**coClustering**] and Bipartite Modularity Optimisation. Since the majority of community detection algorithms are for unipartite graphs, we perform a projection into a one-mode network [**bipartiteProjection**]. 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 [**guimera2007module**].

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 [**6876043**]) 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

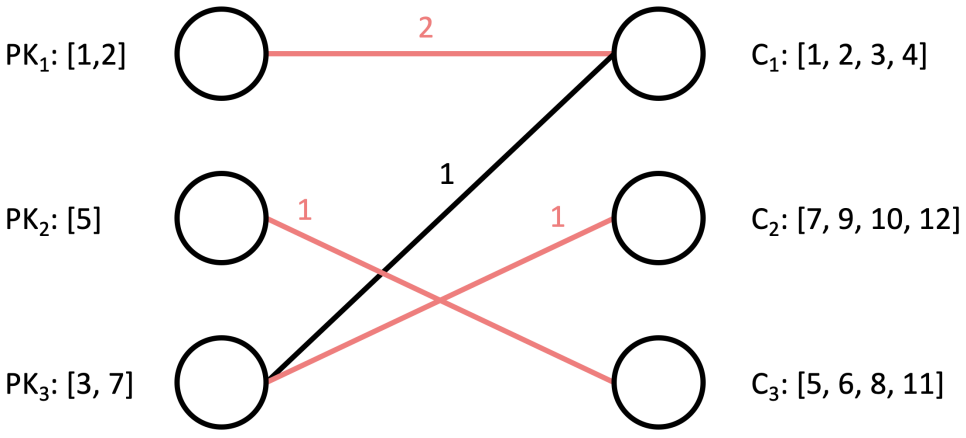
likely to be for the user.

4.3.4 . 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, 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. [**Assignment**].

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), the edit distance.



r0.5

Red edges represent the prior knowledge matching For example, given a clustering of 12 nodes $N = 1, 2, \ldots, 12$, 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 subsection 4.3.4. The edges of the matching correspond to the matching between the PK-groups and the clusters. The edit distance is then equal to the sum of all the weights of the bipartite graph minus the sum of the weights of the maximum matching (in red), thus equaling $5 - 4 = 1$. In other words, we only have to move the node 3 from PK_3 to PK_1, for every PK-group to be a subset of an unique cluster, with no overlap.

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

4.3.5 . Ranking the Algorithms

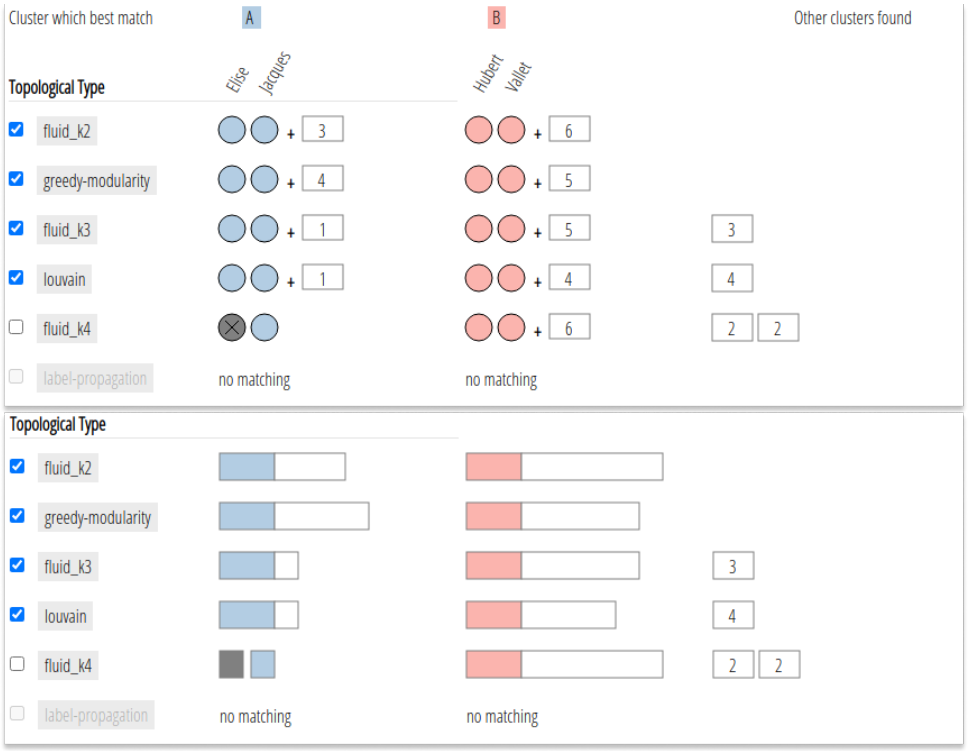
The algorithms are ranked by their degree of matching with the prior knowledge, using the edit distance. We also introduce a *parsimony* criterion if there is a tie between two or more algorithms. The algorithm with the smaller number of other clusters will be shown first, as the results are easier to 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.

4.3.6 . 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. subsection 4.3.6 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 **??**) :



r0.5

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.

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 subsection 4.3.6 we first see the cluster which best matches the blue PK-group, and then the cluster which best matches the red PK-group. In each cluster we see colored dots for each person that matches, and dark gray dots with a X for no match. Additional nodes in the cluster are represented as white dots with a number next to it. On the right most we see how many other clusters (if any) have been found by the algorithm - also represented as white dots with a number next to it.

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

The alternate modality of representing the matches (shown at the bottom of subsection 4.3.6) 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* (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.

4.3.7 . 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 view (see Figure 4.2). In this view, each line corresponds to a person in the graph, and each vertical line represents an hyperedge connecting them [**paohvis**], in a way visually similar to the UpSet representation [**lex2014upset**] but semantically different. Alternative graph representations are available as well—such as node link diagrams—but the 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 view (see Figure 4.2 and also **??** for overall context). By default, the 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 4.2 PK_louvain is selected as the base algorithm for the grouping of names in the list. We see that there is very good consensus on the red cluster, but in the blue cluster only 4 out of 7 algorithms see Joseph as belonging to it. Others see him as belonging to the red cluster. In *Others*, 4 algorithms consistently disagree by assigning 3 more nodes to the blue cluster. There are clearly many ways to cluster data, and users must decide the more meaningful one, based on their deep knowledge of the people in the network before validating clusters, possibly by re-reading source documents or gathering more.

Consolidating the prior knowledge clusters

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

Partial Copy. By clicking on one of the icons or dragging the cursor down on a set of icons, users validate the suggestion(s) of an algorithm, adding colored squares in the consolidation column. Once this validation is done, the squares do not change color anymore and represent the user's final decision (unless changed manually again). Figure 4.3 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

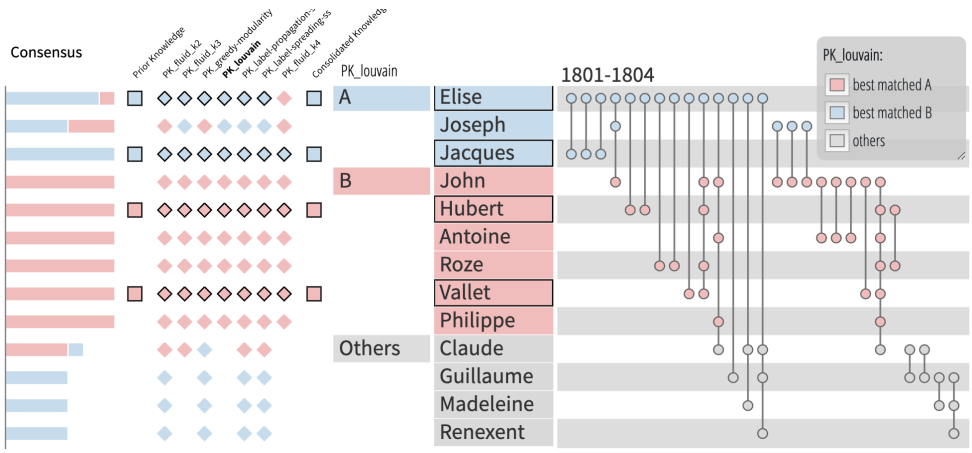


Figure 4.2 – Reviewing and comparing results of multiple algorithms. One algorithm is selected to order the names and group them, but icons show how other algorithms cluster the nodes differently, summarized in the consensus bar on the left.

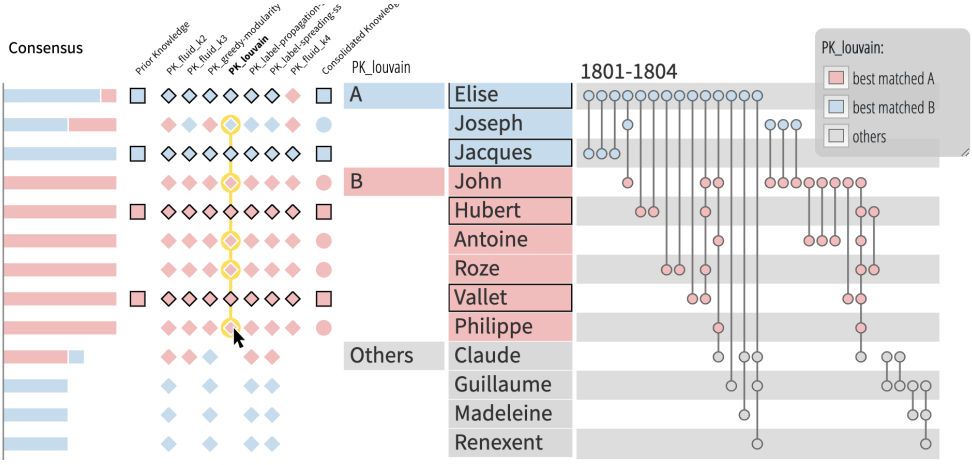


Figure 4.3 – The user quickly drags on consecutive icons (in yellow) representing the suggestions made by one algorithm to validate node clustering. Once the cursor is released the validated nodes appear as squares icons in the Consolidated Knowledge column.

dragging over the icons. When done the status of the nodes in the Consolidated Knowledge column (rightmost) will change to square.

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

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

Consolidating extra clusters

The last step of PK-clustering aims to find new clusters for the nodes that have not been validated yet, based on the consensus of the selected algorithms. The suggestions are made from the point of view of one clustering algorithm that the user can change along the process. First, the user selects one algorithm in the PAOHVis view and the nodes are grouped by the clusters found by the algorithm. 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 subsection 4.3.4) . Once the matching is done, the consensus of one cluster is computed as the sum of the cardinalities of the intersections between the cluster and all the other clusters of the other algorithms matched with it, divided by the number of nodes of the cluster.

When users hover over one cluster name, a new color is given to that cluster (green) and new (green) diamonds appear for each algorithm that match the cluster and for each node that is assigned to the cluster (Figure 4.4). Users can therefore see if the selected cluster is consensual, and with which algorithms. The top part of Figure 4.4 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

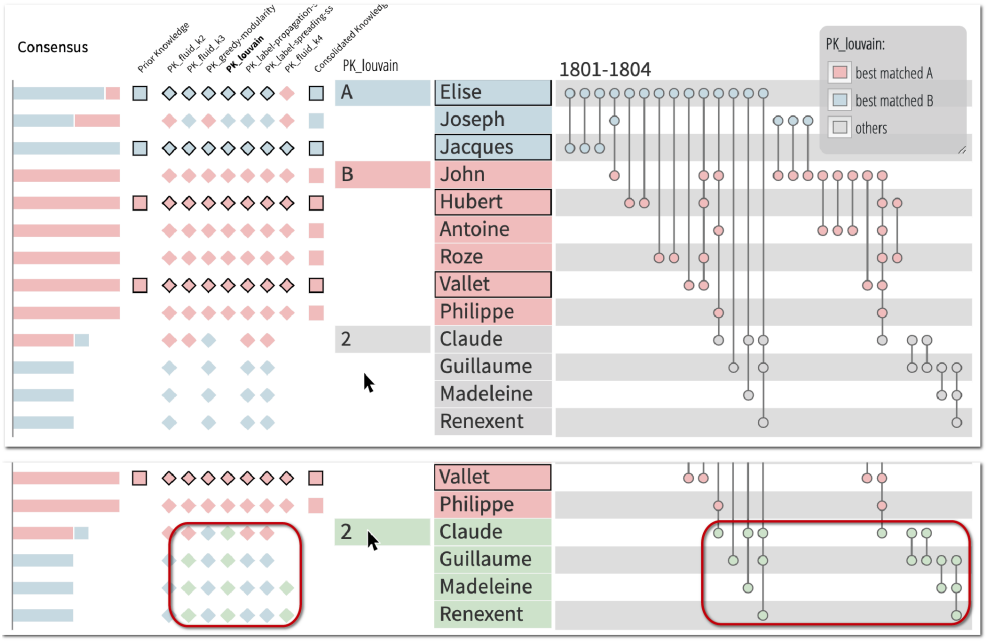


Figure 4.4 – Suggestion of extra clusters. The two PK-groups (red and blue) are validated (nodes in the consensus column are all squared). One extra clusters is proposed by the Louvain algorithm, labeled as 2. Hovering over the cluster 2, the consensus is displayed by the green diamonds. This feedback is also visible in the graph.

association proposed by the current algorithm, they can validate it by clicking on the cluster name. This will create a new group, so the user can classify the nodes into this new group, as seen before (subsubsection 4.3.6) : 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 **??**.

4.3.8 . 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.

4.4 . 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.

Figure 4.5 – The user has multiple possibilities to compare the consensus. Looking at *Claude* node, four algorithms suggest the red cluster, one algorithm the blue cluster and one the green one.we can remove this figure if needed

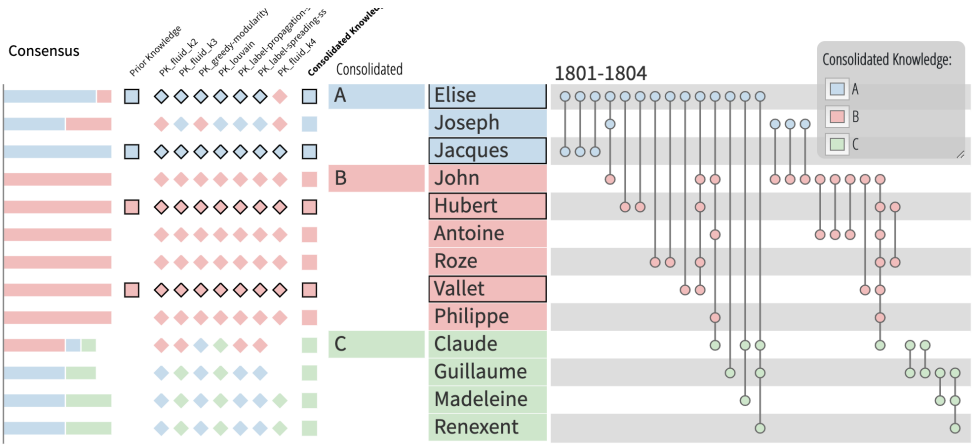


Figure 4.6 – The dataset has been fully consolidated. The persons are grouped and colored by the consolidated knowledge. The user decided to assign Claude, Guillaume, Madeleine and Renexent to cluster C, by taking into account the graph and the consensus of the algorithms.

4.4.1 . Marie Boucher Social Network

We asked our historian colleague her prior knowledge on her network about the trades of Marie Boucher [**Dufournaud17**], 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, An-

Figure 4.7 – Ranking of the clustering algorithms based on their matching with the PK for the Marie Boucher dataset

theaume, 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. For the Marie Boucher case study, the report would write check : *"Nine out of thirteen algorithms matched our prior knowledge for three communities, supporting our hypothesis and assigning minors actors to the clusters. Furthermore, the clusters were strongly correlated with three time periods, consistent with the life history of Marie Boucher and the constitution of her trading network."*

4.4.2 . 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 [**infovisfaq**] 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 [**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 (**??**) 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 4.8 – Final consolidated partition of the Marie Boucher dataset. The persons are colored and grouped by the consolidated knowledge. We can see that the final grouping makes sense in the graph visualization on the right. Only one person is not part of any group : Jacques Souchay.

Figure 4.9 – Computing the Lineages of VAST authors : Prior Knowledge from Alice and results of the clusterings matching it.

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 (**??**). The history of VAST in Europe, very much centered around D. Keim and the VisMaster project [**VisMaster**], 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.

4.4.3 . 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.

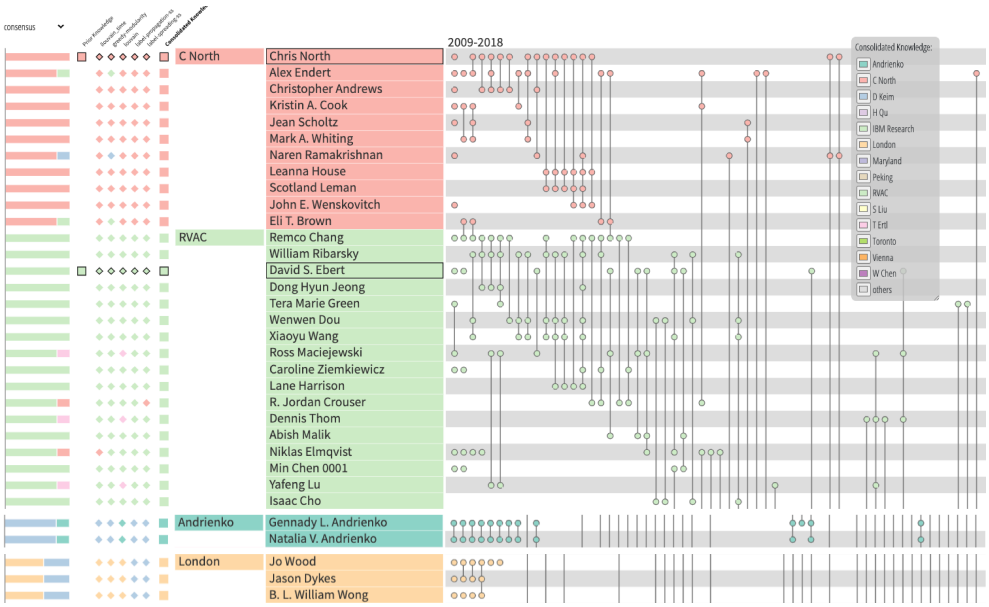


Figure 4.10 – Four consolidated groups in the VAST dataset : C North, RVAC, Andrienko and London

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, 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.

4.5 . Discussion

As presented in subsection 4.2.6, 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.

4.5.1 . 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

4.5.2 . 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 **??**, 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 **??** 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 (50–500 nodes).

In order to improve the computational scalability, we will implement progressive techniques to deal with larger sizes [**Progressive**]. 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 (selecting one algorithm based on the ranking, or using the consensus slider to consolidate all the nodes at once). Pixel-oriented visualizations [**keim2000pixel**] would facilitate the review of consensus for a large number nodes and clusters. Classic techniques like zooming or fisheye views [**Jakobsen06-fisheye**, **rao94**] would help as long as names remain readable, which is critical to our users.

4.6 . 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.

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

5.1 . Introduction

Social historians and sociologists aim at retrieving and studying facts about a specific region and period of time that they focus on. Their work essentially relies on documents—such as marriage acts, census records, surveys, and business contracts—to gather information about the life of important actors that they explore in-depth, or to draw conclusions on social aspects of groups in the society of that period and place. Instead of drawing conclusions from their gathered knowledge and interpretations of the documents, a more systematic approach consists in constructing a social network from the documents and following a Social Network Analysis (SNA) approach [**wetherell_historical_1998**]. For this, they need to encode their documents to extract the persons and any other useful information in the text and transfer it into a structured file or a database. Social scientists can then explore, validate, or refute their hypotheses by observing and analyzing the network structure and the connectivity patterns between the entities of the resulting network. They also want to visually explore their data to generate new insights and hypotheses.

Currently, social scientists often model their datasets as simple networks where the nodes are the persons mentioned in the documents. 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 [**grandjean_analisi_2017**, **rossi_exploration_2014**, **shafie_hypergraph_2017**].

In addition, historical documents include time and geospatial information corresponding to the date and location of the events they refer to. Documents often mention additional information on the persons, such as their sex, profession, and date of birth. These are often essential to understanding underlying social phenomena, as time, space, and social status play an important role in social dynamics. For these reasons, historical sources and the underlying social phenomena they refer to can be modeled well by *bipartite* with *roles*, *multivariate dynamic* networks. *Bipartite* means that both persons and documents (or events, that are often witnessed by physical documents) are modeled as typed nodes. *Multivariate* means that the nodes and links can carry additional attributes. *Dynamic* means that time is a

mandatory attribute of documents. Furthermore, a link created between a person's node and a document's node (when the person is mentioned in the document), has an associated link type that models the *role* of the person in the document/event. Additionally, documents can optionally carry a geographical location. This model unifies several social network models and allows to model the historical sources with any transformation, simplification, or loss of information [**cristofoli_aux_2008**].

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

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

we thus focus on three high-level tasks : exploration, queries, and comparison of this type of network. Users can explore the data using two layouts : a node-link bipartite view showing the sociological structure of the network, and a map layout based on the geolocation of documents. We designed and implemented a new visual graph query system that allows us to build both topological and attribute constraints, based respectively on a node-link interactive representation, and dynamic widgets. For this, we rely on the Neo4j graph database [**neo4j**] and its language *Cypher*. Most visualization systems offer dynamic queries to hide the complexity of query languages. However, using a rich data model, some queries are much easier to refine using scripting than dynamic queries. We implemented dynamic queries that also show the translated Cypher queries, and inversely, can translate textual queries into visual queries. With that interface, social scientists can start building their queries with simple widgets and, if needed, complement them by editing the query, alone or with the help of power users. On top of that, they can easily copy and paste the textual query to share the current state of the query and associated results with someone else or to start an analysis session from a previous result. also implements subgraph comparison techniques, allowing the comparison of networks, network-related measures, and attribute distributions between the entities returned by the queries. We validate the query and comparison system with a usability study and we demonstrate can be used to answer sociological questions by describing in depth several real-world use cases.

After the related work section, we describe our data model in detail using

four use cases, and present our system , with the design of the visual query and comparison features. Finally, we present two use cases demonstrating the utility of our system, showing it can be used to explore the complex historical data and allowing them to answer several of their questions using queries and comparisons. Our contributions are :

— The design and implementation of a graph query system, synchronizing the visual representation of the query and the associated script ;

— The design and implementation of visualization and interaction techniques aimed at comparing subgraphs, in terms of topology, attributes, time, and geographical location.

— A usability study and two real-world use cases demonstrate the utility of the system to answer socio-historical questions.

5.2 . Related Work

Social networks have been studied from the perspective of SNA, network visualization, graph databases, and visual analytics.

5.2.1 . Modeling Social Networks

SNA aims at addressing sociological questions using mathematical methods based on graphs. It started by encoding social networks using standard graphs [**freeman2004development**], defined as $G = (V, E)$ with V a set of vertices and $E \subseteq V^2$ a set of edges. Social entities become vertices, usually called *nodes*, and relations become edges, usually called *links*. The SNA graph models have evolved to better take into account real properties of social networks, such as types of actors using labeled graphs, importance of actors or relations with weighted graphs, bipartite graphs later to represent networks where relations only exist between two types of node, such as organization and employees where the relations link employees to organizations but not employees to employees or organizations to organizations. They have been generalized as multi-mode graphs when more than two types of nodes exist, such as documents citing persons and places.

Temporal (also called *dynamic* and *longitudinal*) graphs are also important in SNA, with multiple models to associate time with vertices, edges, or to graph snapshots $G_1, G_2, \ldots, G_n$ at time $t_1, t_2, \ldots, t_n$, each graph sharing the same vertices [**STARDynGraphs**].

Multivariate networks, i.e., graphs where vertices and edges can be assigned multiple "properties" or "attributes", possibly with an associated value, are less used in SNA. These attributes are often considered secondary, the emphasis of SNA being on the graph topology, its features, measures, and evolution.

In contrast to SNA, *Graph databases* are concerned with the concrete encoding of networks in computer files and memory and less by their mathematical analysis. General graph databases and RDF graphs are popular network data models. To

be used for SNA, a network data model needs to be transformed into a suitable graph through a query language. Popular graph databases such as Neo4j [**neo4j**] can be queried and modified using domain-specific languages (called Cypher for Neo4j). Graph databases can model any kind of complex social network, but they do not provide guidance for modeling real social networks. To be used, they require to understand graph data modeling, and very few social scientists have that skill. Additionally, their flexibility in data modeling produces fragmentation of data models used in real applications, hampering the development of interactive visual interfaces that rely on specific data models.

Modeling Social networks can also be done using the *semantic web*, which relies on a graph representation called the "Resource Description Framework" (RDF). RDF models any graph and its attributes as a set of triples of the form *subject*, *predicate*, *object*. This representation can model a large variety of graphs, including social networks. However, like graph databases, it does not enforce one particular modeling and produces an important fragmentation of data models, which is detrimental to the design of graphical interactive tools. Even if the standard "Friend of a Friend" (FOAF) [**FOAF2007**] RDF data model is designed for social networks, its use is complex and leaves a lot of freedom for modeling concrete social networks. Like graph databases, social scientists are rarely able to model their data with RDF and even less to use the standard semantic web querying language SPARQL [**sparql**] ; they need higher-level tools.

Historians, demographers, sociologists, and anthropologists have been designing specific data models for their social graphs, based on genealogy or more generally kinship [**hamberger:halshs-00658667**]. For genealogy, the standard GEDCOM [**gedcom**] format models a genealogical graph as a bipartite graph with two types of vertices : individuals and families. Both types can have attributes and be associated to *events*, such as birth, death, marriage, and many more. These events are dated. A genealogy graph is therefore encoded as a bipartite, multivariate, dynamic, social network with roles since individuals play particular roles in families. The PUCK software [**PUCK**] has extended the GEDCOM format to adapt to more flexible kinds of family structures for anthropology studies ; it has also extended the types of relations and events to handle different kinds of ties between individuals or families. Due to its evolution, it is complex to understand but reflects the needs of social scientists.

5.2.2 . Visual Analytics for Social Networks

In addition to mathematical analysis and data modeling, several systems have been developed to support the visual exploration of social network. They all use visualization but not all of them are interactive. Most of the SNA-oriented systems such as Gephi [**Gephi**], Pajek [**batagelj_pajek_nodate**], UCINET [**ucinet**], and NodeXL [**NodeXL**] visualize networks with node-link diagrams, using several types of layouts. More recent systems use alternative representations such as matrices, TimeArcs, and geographical maps for geolocated entities, and others [**vistorian**,

valdivia:hal-02264960]. More specialized systems focus on visualizing bipartite networks : GeneaQuilts [**GeneaQuilts**] visualizes genealogical graphs as bipartite graphs with a cascade of matrices. However, that visualization is very specific to genealogies, although it has been extended by PUCK to visualize other bipartite structures.

Jigsaw [**DBLP:journals/ivs/GorgLS14**] is a visual analytics system based on documents, designed for intelligence analysis. Its data model is based on a collection of time-stamped documents. Each document contains text and identifies *named entities* such as persons, places, organizations, etc. Jigsaw provides a large set of visualization techniques to explore the documents in detail or as overviews, such as a node-link diagram of entities (a multi-mode network), a list view, and many more. Compared to PUCK, Jigsaw does not handle the *roles* of people in documents, it merely considers each mention of a person as a link with no finer precision.

PAOHVis [**valdivia:hal-02264960**] visualizes hypergraphs of documents and persons as dynamic hypergraphs. A hypergraph is a generalization of a graph where (hyper) edges can contain one or more vertices instead of exactly two. Internally, a hypergraph is represented as a bipartite graph with a document vertex connected to person vertices. Therefore, it also encodes social networks as bipartite dynamic social networks. It also supports *roles* (link types) and node types or groups. Yet, PAOHVis does not handle vertex or edge attributes.

5.2.3 . Visual Graph Querying

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

Finding and extracting a subgraph of interest in a bigger graph is an old problem in SNA. Constructing and querying a pattern from a graph requires knowledge of graph databases and query languages. To lower the complexity barrier, several visual graph query systems have been developed to allow analysts to rapidly build and refine their queries visually. GRAPHITE [**chau_graphite_2008**] and VERTIGo [**cuenca_vertigo_2021**] allow specifying a graph query as a node-link diagram that the user creates interactively. Shadoan and Weaver [**shadoan_visual_2013**] use a similar concept with hypergraphs to filter multidimensional data. Other systems such as VIGOR [**pienta_vigor_2018**] only visualize the query after it has been written using a script language. However, these visual systems are limited to topological queries including constraints on the vertex and edge types, they do not support constraints related to general attributes and time associated with vertices and edges.

5.2.4 . Visual Graph Comparison

Gleicher et al. [**gleicher_visual_2011**] propose a taxonomy of visual comparison designs of complex objects. They claim any comparison system can be classified

into one (or a mix) of the three following categories : (a) juxtaposition, (b) superposition, or (c) explicit design. Yet, few systems support comparison tasks on social networks.

Andrews at al. [**andrews_visual_2009**] describe a technique to compare several graphs, using a combination of juxtaposition and superposition techniques. The two candidate graphs are shown side by side, along with a third view composed of a fusion graph highlighting both the shared nodes along with the non-shared nodes with different colors. Freire at al [**ManyNets**] 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 [**HascoetD12**] describe a system to match and compare graphs using superposition, focusing on the topology, not taking into account attributes or time. Tovanich at al.[**tovanich_vast_2021**] propose a visual analytics tool to compare multivariate, sometimes bipartite, dynamic graphs and find common structures. Yet, their tool does not handle attributes or roles, and is designed for the specific task of matching a subgraph to a large graph.

5.3 . Task Analysis and Design Process

We designed our tool in collaboration with historians who have historical documents data that fit well our bipartite multivariate dynamic networkmodel. We first collected all the questions they had on their data and what they wanted to see in a visual interface. By analyzing the questions we leveraged tasks and requirements. We designed the interface from the requirements with continuous discussions with our collaborators.

5.3.1 . Use Cases

We describe here four example projects coming from close collaborations, involving regular meetings and multiple interviews over two years ; from these collaborations emerged our proposed network model. All these datasets are textual corpora constituted of historical documents mentioning people with complex relationships. They are thus well modeled by bipartite multivariate dynamic network. We also list the main questions our collaborators had and the graph queries extracting the information to start answering them. The full answer involves visualizations of the query results that we describe in the next section.

1. Analysis of the social dynamics from **construction contracts in Italy in the 18th century (141 documents, 272 persons) [Cristofoli2018].** The corpus is made of contracts for different types of constructions in the Piedmont area in Italy. People are mentioned in three different roles : *Associates* (S) who participate in the construction, *Guarantors* (G) who bring financial

Main Tasks	Subtasks	Views	Constraints
Bipartite Graph Exploration	T1.1 Overview of the network	V1	A node-link represent The geolocation of ev
	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 e
	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 5.1 – Tasks to support during exploration, according to our expert collaborators, split into 3 main high level tasks.

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

Question 1 Do approvers act as bridges compared to associates and guarantors ?

Query 1.1 Request all approvers occurrences

Query 1.2 Request all associates and guarantors occurrences

Question 2 What are the differences between Turin (Torino) and Torino close area according the contracts ?

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

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

Question 3 Who are the persons of the extended Zo family

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 term of contracts and activities

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 ?

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

Question 6 Are there people mutually guarantor to each other in different contracts ?

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

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

Question 7 See the trajectory of life for an individual (birth, living, marriage, death)

Query 7.1 Select one person, and all his documents (to extract the geolocated places)

Question 8 See the trajectory of life for a family

Query 8.1 Select one birth certificate with the child and parentsNot clear

Question 9 What are the main migrations ?

Query 9.1 Select the persons with a geolocated birth certificate and death certificate

Question 10 Is there differences between the migrations in the 18th and 19th centuries ?

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

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

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

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

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

Query 12.1 Select all persons in the first half of the 19th century who have a profession mentioned

Query 12.2 Select all persons in the second half of the 19th century who have 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) [moutoukias2016buenos].** 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 ?

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

Query 13.2 Select marriages with spouses and witnesses, where the spouse and witness have the same grand parents

Question 14 Who are the persons with 2 marriages with certain amount of delay ?

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 coming from the persons marrying in Buenos Aires ?

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

4. Socio-political analysis of **migration of ethnic Germans from communist Romania to West Germany in the 20th century (ongoing work) [diminescu:hal-02556007].** 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 date of birth.

Question 16 What member of their family emigrant often join ?

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

Question 17 What price had to pay the emigrant, given their socio-economic profiles ?

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

5.3.2 . Tasks Analysis

Most of the questions we collected from our collaborators could be answer by isolating a subgroup of entities and analyzing them in context on the whole network, or by comparing two subgraphs, in term 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 in three main parts : 1) Exploration of the data, 2) Queries, and 3) Comparisons. The tasks are described here and summarized in **??** :

1. **Exploration of bipartite multivariate dynamic network**. The visual interface must allows exploration of this specific type of graph, 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.

Use Case	Id	Question	Queries
Piedmont Constructions (#1)	1	Do approvers act as bridges compared to associates and guarantors?	Request all approbator occurrences
			Request all associate and guarantor occurrences
	2	How Torino and Torino close area vary according their contracts?	Request all documents located in Torino, with the persons mentioned
			Request all documents located in Torino area, with the persons mentioned
	3	Who are the persons of the extended Zo family	Request the persons of the Menafoglio family and their N+2 ego network
	4	Compare the Menafoglio and Zo families in term of contracts and activities	Request all Menafoglio people and their document
			Request all Zo people and their document
	5	Who are the persons having the 3 roles?	Select persons with one associate, guarantor and approbator roles in 3 different documents.
	6	Are there people mutually guarantor to each other in different contracts?	Select pairs of people connected each to the two same document, with a guarantor and an any link.
French Genealogy (#2)	7	See the trajectory of life for an individual (birth, living, marriage, death)	Select one person, and all his documents
	8	See the trajectory of life for a family	Select one birth certificate with the child and parents
	9	What are the main migrations	Select the persons with birth certificates and death certificates which are both geolocated.
	10	Is there migration difference between the 18 and 19 centuries	Select the persons with birth certificates and death certificates which are both geolocated and from the 18th century.
			Select the persons with birth certificates and death certificates which are both geolocated.and from the 19th century
	11	In Haute-Vienne and maybe Cote d'Armor, is there cycle of living (in cities), every 10/20 years	Select persons with their census documents located in Cote d'Armor and Haute-Vienne
	12	In 19th century, was there a decrease in social status and professions?	Select all persons in the first half of the 19th century who have a profession mentioned
			Select all persons in the second half of the 19th century who have a profession mentioned
Marriages in Buenos Aires (#3)	13	How are spouses and witnesses linked in their family network	Select marriages with spouses and witnesses, where the spouse and witness have the same parents
			Select marriages with spouses and witnesses, where the spouse and witness have the same grand parents
	14	Who are the persons with 2 marriages with certain amount of delay	Select persons in 2 marriages as husband or wife. Put a constraint on the difference of time in the marriages.
	15	Where are coming from the persons marrying in Buenos Aires	Select persons with a birth certificate located not in Buenos Aires
Migrations from Romania (#4)	16	What member of their family emigrant often join?	Select all migration documents with the emigrant and the person they are joining.
	17	What price had to pay the emigrant, given their socio-economic profiles	Select people who are mentioned in a budget document and a migration document.

Table 5.2 – Most important questions our four collaborators shared with us on their respective datasets. We provide the original questions and the associated queries which can be used to answer the questions. Questions with two queries are answered by comparing the results of the two queries.

2. **Applying filters**. To answer their questions, users need to be able to apply filters on 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).

5.4 . The System

is designed to visualize, explore, and analyze social networks encoded as bipartite multivariate dynamic network. It dynamically collects the node types, roles, sub-types, and attributes when reading the network from the database. is constituted of four main panels, split in different views as shown in **??** : the query and comparison panel, the bipartite graph visualization panel, the map visualization panel and the query results panel.

5.4.1 . Visualizations

presents a social network with multiple visualizations highlighting different aspects of the data. The visualizations are linked when it makes sense.

V1 : Bipartite Node-Link Diagram The bipartite node-link visualization panel shows the network using a force-directed layout. Node-link representations are very common in social sciences [**Gephi**] [**batagelj_pajek_nodate**] and it was 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 model the mention of a person in a document. This view provides an overview of the data by showing the structure of the network (T1.1) and the roles of the persons in their different documents (T1.2). Attribute values can be overlayed on the nodes using colors when users select an attribute. It allows to detect patterns relative to attributes, in context of the topology of the network (T1.2, T1.4, T1.5). The view allows pan & zoom, and selection for a good navigation.

V2 : Map View The map visualization panel on the right shows an event-centric view, displaying only the geolocalized event nodes on a map. By default, only

event nodes are shown. The system can create links between event nodes which share persons with a threshold given by the users. Persons are not directly shown in this view as they do not have a unique location. This map view presents a transformation of the bipartite graph, focused on the geospatial information that is very important to the social scientists (T1.3).

As we collaborate with historians who study different periods, we can not use recent map backgrounds such as the ones provided by OpenStreetMap. We therefore provide a map background with only these non-administrative features : elevation, lakes, rivers, types of environment. We also show the most famous cities as the majority of them existed in the past and to give points of reference. The map use Natural Earth tiles and vector data.

The two views are coordinated and selecting or hovering over an event node in the bipartite view highlights it in the map and vice versa, while hovering a person node highlights all its corresponding documents in the map, rapidly showing the events' location of this person.

V3 : Entities Tables All the persons and the documents of the loaded dataset are listed in two separate tables, showing the attributes of the entities. This way users can order the entities according 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.

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

V5 : Attributes View All the attributes in the network are shown as buttons in the bottom right of the interface, sorted by their associated node type (person, document, and both). They can be quickly visualized by hovering over the button, producing two effects : it colors all the nodes on the two views according to their attribute values, and it shows a plot of the distribution of the selected attribute, as shown in **??**. By clicking on the button, the visual encoding and distribution remain selected. 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).

5.4.2 . Query Panel

The query panel allows to rapidly build queries visually, with both topological and attribute constraints. The visualization of the query is synchronized with the Cypher query sent to the database. Modifying one representation will update the other, allowing users to build a query visually and refine it in Cypher when appropriate. In this section, we describe all the features and interactions allowing to build

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

V6 : Node-Link Dynamic Query The interactive node-link diagram allows to build a subgraph query graphically, which represents a topological constraint (T2.1). To find persons who are mutually guarantors, we first create one person and two documents. We link the person node to the first document with a link that is not typed, as in **??** left, and link it the the second document with a Guarantor link, as in **??** middle. We then create a second person node and link it to the two documents with reversed link types.

The query subgraph is built and edited interactively. At each modification, the subgraph is converted into a Cypher query, run in the database, and all its matches are returned and highlighted in the main visualizations. Three modes of interaction are available through the top-right menu : *selection, addition*, and *deletion*. The *selection* mode allows to drag the nodes in the panel, while the *addition* and *deletion* modes allow the following actions :

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

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

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

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

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

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

Users build concrete subgraphs with the same representation as in the bipartite graph view : a visual query is a graph template. Each role (link type) is rendered using a color. We can also create untyped links using the *Any* value, which will be matched by all the existing link types. We also allow creating links that can be matched by several selected link types in the graph, by checking several possible types for one link. These links are represented by a dashed line with the colors of the possible types. All possibilities for link creation are presented in **??**. Note that when a node and link is created in the query, it is given an identifier starting with $pers$ for a person, doc for a document, l 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.

[width=0.6]static/figures/ComBiNet/OriginalPaperFigures/links/allLinks.pdf

Figure 5.1 – All link creation possibilities : Any link type (left), one selected link type, here guarantor (middle), and union of several link types (right)

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

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

For the categorical and nominal widgets, selecting several values, by checking several checkboxes, will correspond to the union of the filters. The three widget types are shown in **??**.

To answer our collaborator's question 2, we want to filter the documents which are located in Torino. For this, we first select the whole dataset by linking a person and document node with an *any* link. Then, we select the id *doc1* of the document of our visual node-link query, and the *region* attribute. It will initialize a categorical widget including all the values found in the dataset for this attribute with associated checkboxes. We check the region of interest "*1-Turin Ville*" to select all the documents from this region. The first widget of **??** illustrates the created constraint along with the input buttons which allow the creation of new constraints.

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 visual and textual representations are synchronized, meaning that changing one will update the other and update the results in the visualizations.

Query Results Each modification of the query, whether from the node-link dynamic query, the widgets, or the Cypher text boxes, update the two visualization panels (V1, V2), the entities tables (V3), the graph measures view (V5) and the attribute plots (V6). The nodes and links that do not match (are not retrieved by the query) are grayed out in V1 and V2, and are removed from the persons and

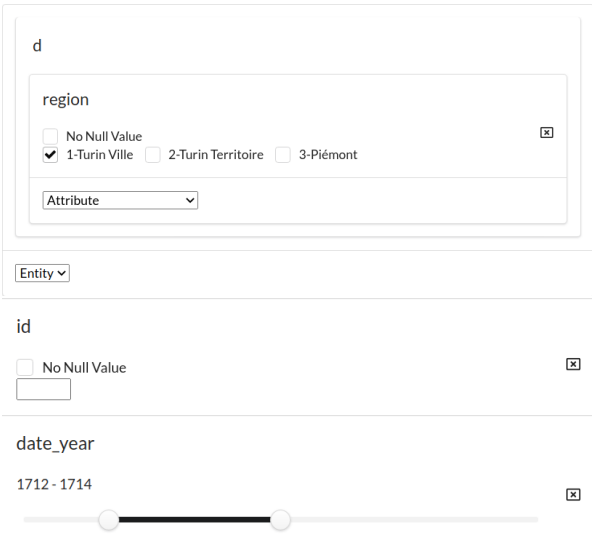


Figure 5.2 – Widgets are associated with the three different attribute types : categorical (top), nominal (middle), numeric (bottom). The categorical widget shows the contextual menu with input buttons to create new widgets on other attributes or other nodes.

documents tables (V3). A third table shows every found occurrence of the created pattern. Users can switch between tables in the table view with tabs. The graph measures are computed on the new graph formed by the union of all patterns found and updated on the graph measures view (V5). Since some measures can be long to compute, the values are computed iteratively and shown in a progressive manner [**fekete2019progressive**] to not block the interface. The distribution plots in the attributes view (V6) are updated, now showing the values of the entities of the new pattern, next to the global distributions.

Attributes Visualization. When users select an attribute in the attributes view (V6), it shows its distribution for the whole network and the query's entities. However, these plots show the aggregated values and we lose the potential value transitions between the nodes of the query. For example, **??** shows a query to list the persons who had a role of "approbator" (noted "A" in green) in a contract after being a "guarantor" (marked "G" in blue) in another contract (using a time constraint). We may want to see if the locations or type 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 for showing the attribute distributions is particularly useful for queries where the nodes have time relationships by definition, such as birth certificates, marriage, or death certificates where we know the order in which

[b]0.30

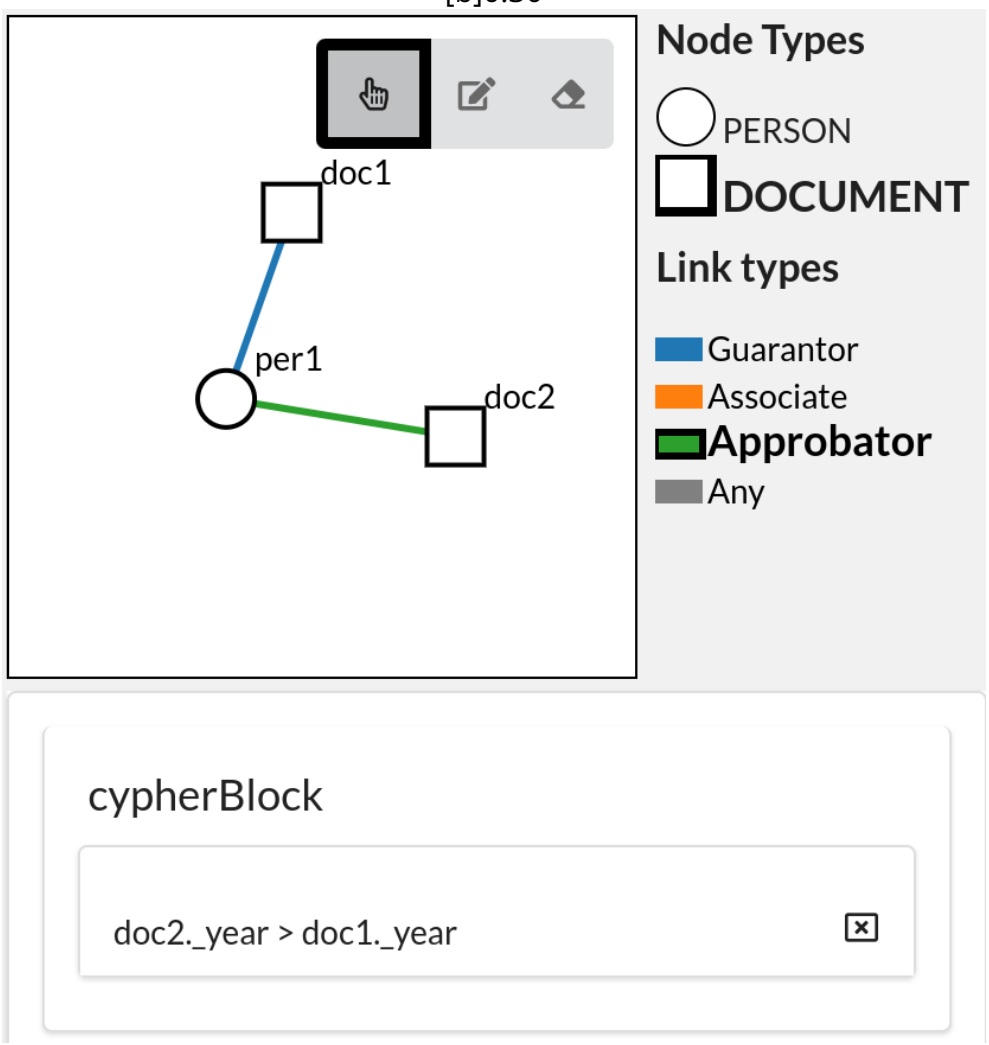
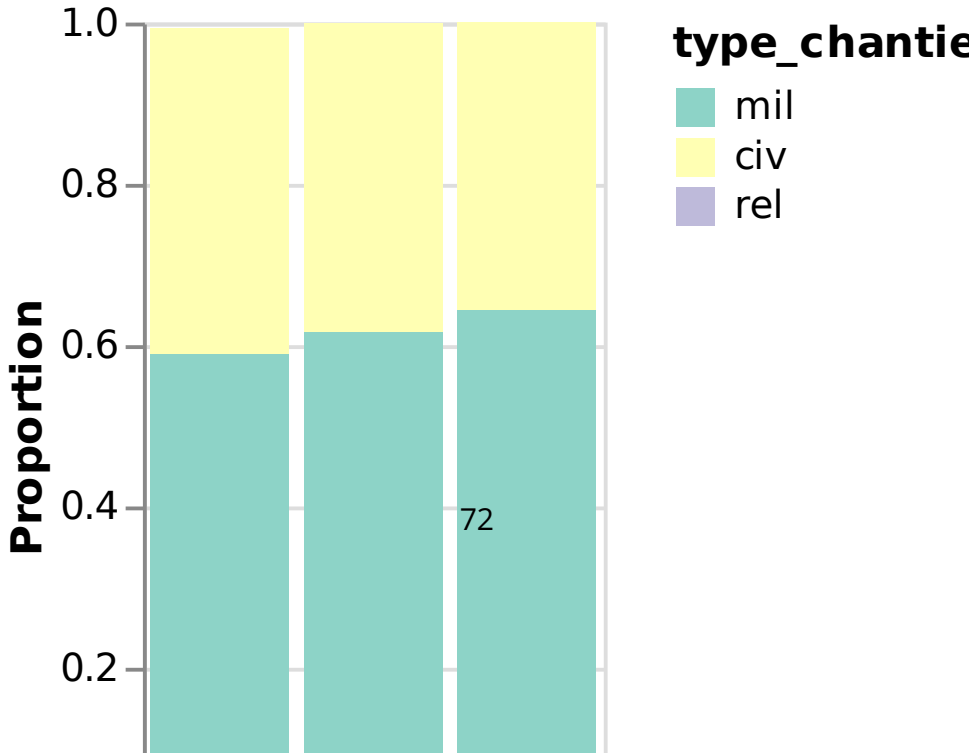


Figure 5.3 –
[b]0.30



these events occurred. It is also useful for queries with user defined time order constraints as in **??**.

V9 : Provenance Tree Each change in the query panel is saved with the computed results, so that the history of the query construction can be shown in the form of a provenance tree (T2.4), managed using the Trrack library [**cutler_trrack_2020**]. Each node of the tree represents a change in the query, with a description label like "New Link" for the creation of a new link. It allows to rapidly visualize the succession of filters applied with their refinements. At any moment, users can click on a node of the tree to go back to a previous query state. It allows reverting to a satisfactory state if a change resulted in disappointing results. Hovering over a node of the tree shows a tooltip with its query visualization. If a change is done after reverting to a previous state, a new branch is created on the tree, allowing to go back to a previous interesting query and refine it in an exploratory way. **??** shows the provenance tree at the bottom of the query panel.

5.4.3 . Comparison

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

To answer several of the questions raised by our collaborators, we need to compare two subsets of the network. For the question 2 from **??**, we want to compare the works in *Turin* with the ones in *Turin Territoire*. Since we previously constructed the query returning all the contracts from *Turin* with the mentioned people, we can return to this point, change the constraint of the *region* attribute from *Turin* to *Turin Territoire* using the checkbox to get the two queries we want to compare. They are shown in **??**.

Topological Comparison In visualization mode, the two graph visualization panels do not change but users can rapidly change between the visual filter of (A) and (B) by hovering over their respective buttons on the comparison menu and thus rapidly comparing the structure of the two resulting subgraphs (T3.1). Similarly, different Boolean comparison operations are available through hovering their respective buttons (shown in **??**-C), such as the intersection, union, and differences of the two filters. Moreover, the summary tab on top of **??**-D allows comparing the different graph measures of the two subgraphs by showing them side by side in a table layout (T3.3). Comparing these measures, such as the number of matched documents or the bipartite density is interesting information in SNA.

Attribute-Based Comparison The comparison of one or several attribute distributions between (A) and (B) is often more useful to answer the historical questions of our users. In the attribute view of the results panel, hovering or clicking on an at-

0.22

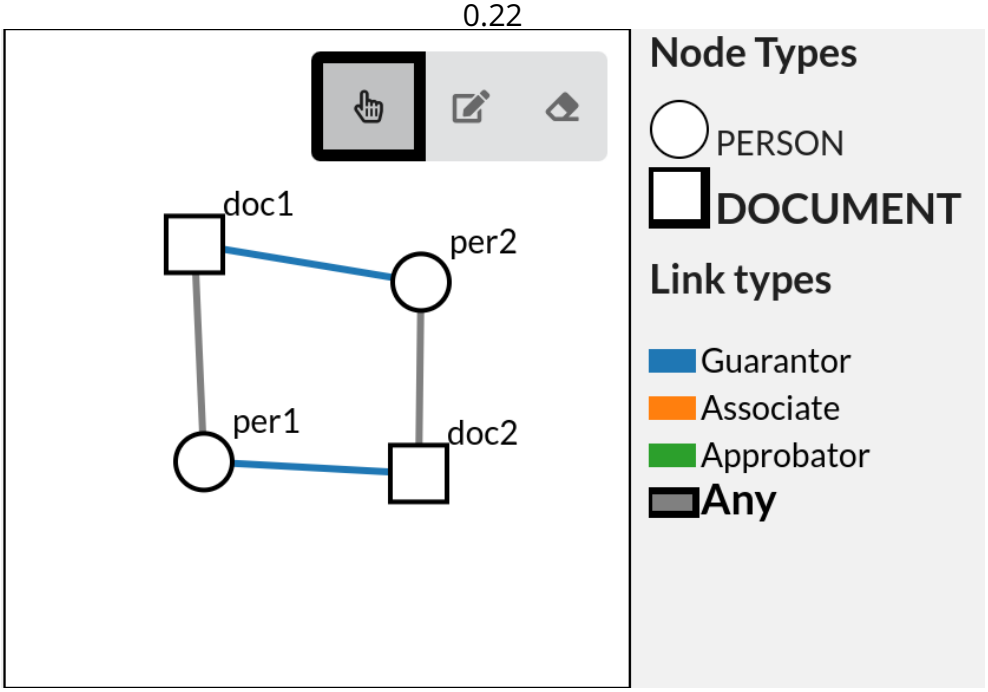
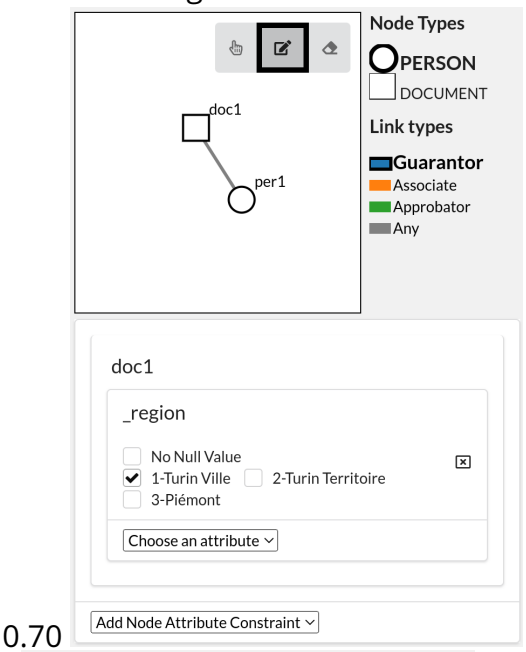
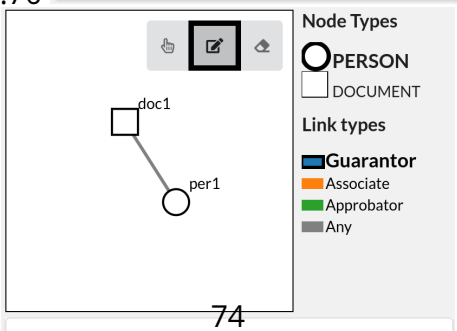


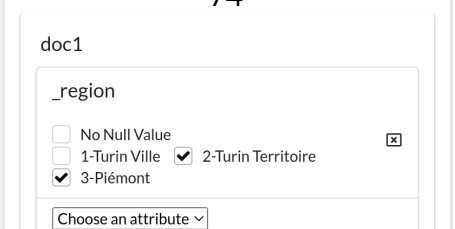
Figure 5.7 –



0.70



74



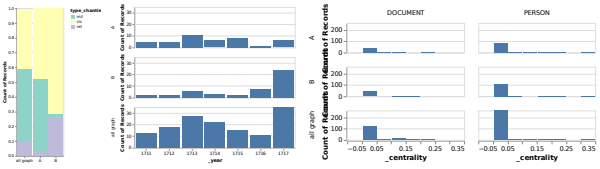


Figure 5.10 – Distribution of two document attributes and one global attribute for the documents and signatories of *Turin* (A), *Turin Territoire* (B) and the whole graph. (top).

tribute name will show the distribution of this attribute in four contexts : the nodes of the whole graph, of the queries (A), (B), and the currently selected Boolean operator (e.g., intersection or union). This allows user 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 close territory. We can also compare the persons who worked in Torino, in Torino's close territory, and in both areas, by selecting the intersection operator. **??** illustrates the comparison charts for different attributes. We can see that the types of construction sites differ between the two regions : the city of Torino clearly have a lot of military sites compared to the territory of Turin which has almost none. This is the reverse for the number of religious sites, which are almost all localized in the territory of Torino. If we now look at the year distribution of the contracts, we can see a difference in the spike of the distributions. The majority of Torino's contracts have occurred around 1713 and around 1717 for Torino's territory. We can also compare the profile of persons who collaborated both at Torino and Torino's territory by selecting the intersection of those two queries. One of the questions the historian had (question 2 of **??**) 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 in 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.

5.4.4 . Implementation

is made of three components : a web visual interface, a python server, and a Neo4j graph database instance. The client interface is written in Javascript, using D3 [**d3**], Vega [**satyanarayan2016vega**], and the Trrrack library [**cutler_trrack_2020**]. The python server is written in Flask, and interacts with the Neo4j instance for query processing. We implemented our Cypher parser with the ANTLR parser generator [**parr1995antlr**].

5.5 . Use Cases

In this section, we describe step by step how our system has been able to specifically answer questions from our users. Since usability issues have not been resolved yet, 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.

5.5.1 . 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 **??**). 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 they 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.

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

Let's now compare the two groups 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 confirm 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 **??** that the Menafoglio family is more specialized in military sites, while the Zo family is doing more civil constructions. 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 in average. This make sense as they act as bridges linking the two families.

5.5.2 . French Genealogy (#2)

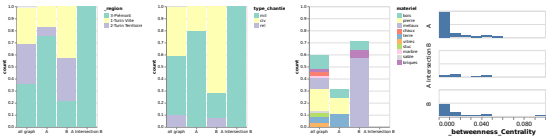


Figure 5.11 – Attributes distributions plots between the whole graph, the *Menafoglio* family (A), the *Zo* family (B), and $A \cap B$, for the *region,type_chantier, materiel type*, and *betweenness centrality*

We describe how allowed to answer one of the main questions of our collaborator of the use case #2. She wanted to detect the largest migrations which occurred in several generations, and in which areas and at what time they happened (question 7 in **??**). The map view shows at a glance (see Figure 3 in the supplementary material) that the majority of events has taken place in three specific regions : one in the west of France, in the top-middle, and in the south-middle.

To find patterns of migrations inside families, we can first make a query representing a simple family by linking a person node to a birth event, itself connected to the parents using a link which is either of type *father* or *mother*. We then repeat the process on 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 **??**). This query returns every person in the graph with their parents and grandparents, along with their respective birth data and death data for the latest person. We also create a constraint on the *department* attribute on all document nodes to specify that we only want to retrieve the events that have an associated location and not a null value, as it is not often the case. 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 returned families. **??** 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. It interested her, so she refined the query by adding an attribute filter on this specific department using a widget. The table view then showed her who these migrant people were and when it occurred. The bipartite visualization panel allows exploring more in-depth this specific group of people.

Afterward, we answered the question 8 (see **??**) of our user closely related to the first one. She wanted to compare the migrations in general between the 18th and 19th centuries. She thinks people really started moving in the 19th century and wanted to confirm it. To answer this, we first created a query to retrieve all 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. We can then compare the two query results using the comparison

0.49

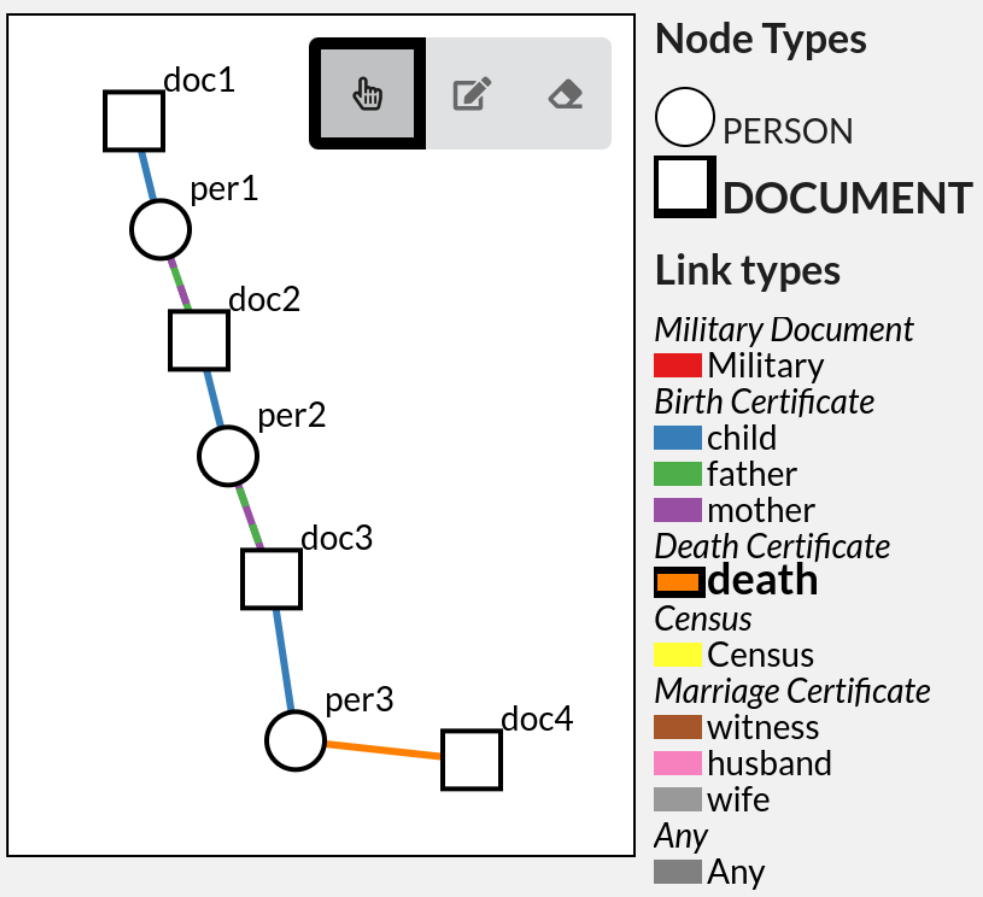


Figure 5.12 – Visual query to find all 3-generation families (documents without department value have been filtered out)

0.49

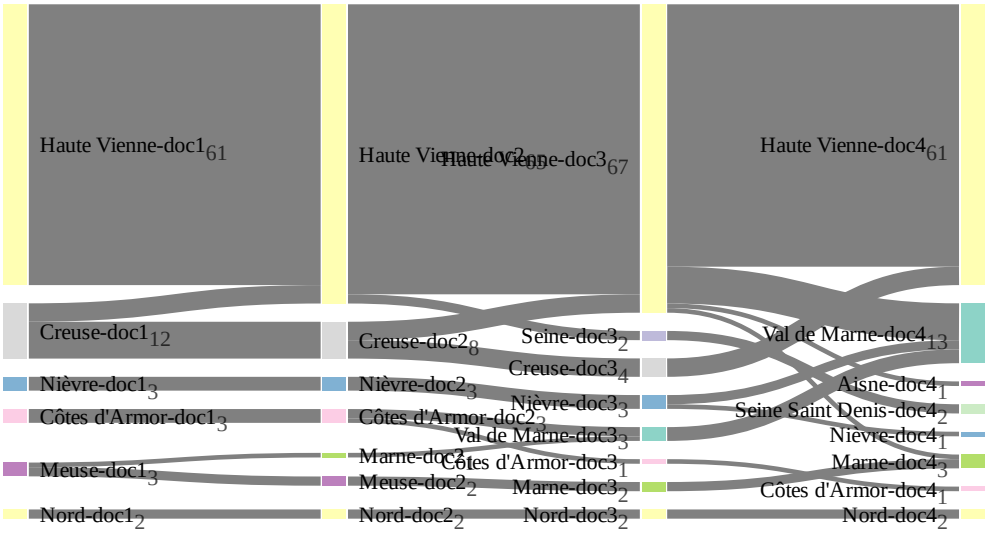


Figure 5.13 – Sankey diagram showing the birthplace of people across generations and the death place

Figure 5.14 – Migrations across departments over three generations

mode, and look side by side at the Sankey graphs related to *departments* as shown in **??**. We can clearly see that people do not move at all in the 18th century, while in the 19th century even if the majority of people stay in the same place from their birth to their death, we can still see more than half move. It thus confirms the hypothesis of our collaborator.

5.6 . Usability Study

Procedure

After showing our tool can be used to answer socio-historical questions, 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 can use it to explore their data, perform queries and comparisons. At each meeting, we asked them to speak aloud, commenting 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 revamp 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.

Feedback

Both historians liked the Sankey view 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 view 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.

Both liked the export capabilities of the system to collect results in CSV or JSON format, as well as the charts in SVG, for inclusion in their publications or for processing with other tools.

5.7 . Discussion

We discuss here several points of potential limitations.

Scalability. We assess the scalability in network size (number of nodes and

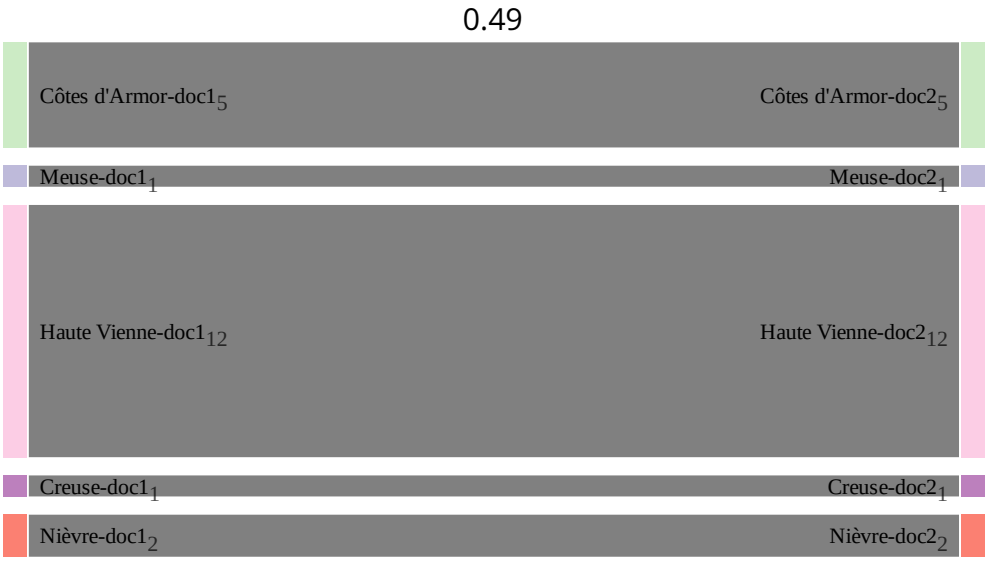


Figure 5.15 – 18th century

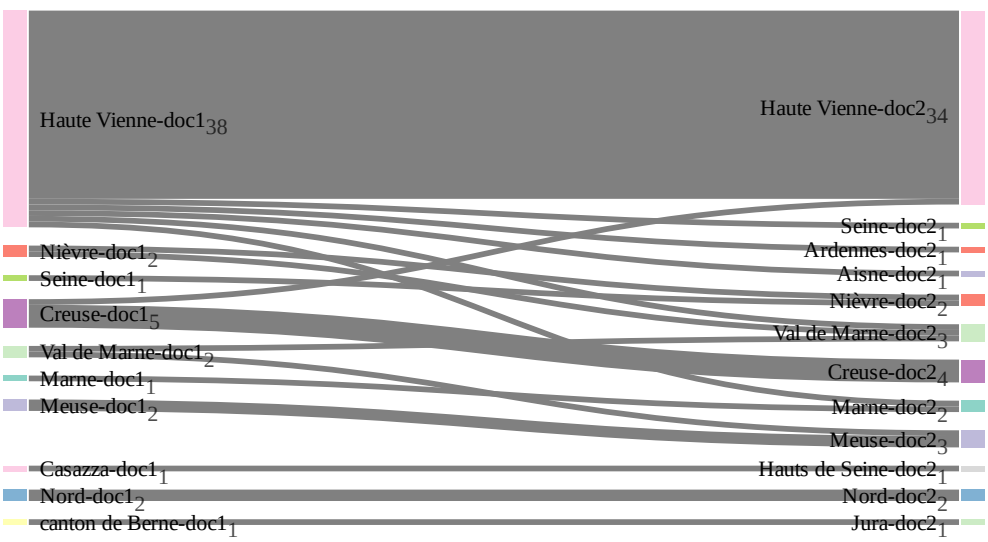


Figure 5.16 – 19th century

Figure 5.17 – Sankey diagrams showing the migration of people in the 18th and 19th centuries, extracted from their birth and death place.

links) with respect of the cluttering and readability of the network visualizations. Our biggest dataset from #3is constituted of 7212 nodes (4886 persons and 2326 events) and 7790 links, after splitting the documents into birth and marriage event nodes. The system allows exploration of network of this size with a decent frame rate. Large networks are usually hard to read due to edge crossings [**shneiderman2006network**]. allows to navigate large graphs with the node-link visualization using zoom & pan and filtering with the query system. It let users focus on subsets of the data, one at the time.

Generalizibility. The system have been designed specifically for bipartite multivariate dynamic networkwhich model well a diversity of historical sources we encountered via our collaborations : marriage acts, birth/death certificates, construction/work contracts, census, migrations forms. However, there exist some special dataset cases where documents can have relationships with other documents, such as letters for example. The model and interface would have to be slightly changed to take into account document-to-document links for these datasets. Moreover, bipartite multivariate dynamic networkcan also be used to model other types of similar data, such as scientific publications or thesis data. Bipartite networks are also used in various other disciplines, such as biology [**klamt2009hypergraphs**] or chemistry [**konstantinova2001application**]. could in theory be extended to these other application domains, by modifying the map view to show other geolocation data related to the entities of the network.

Dynamic representation. The current system allows to explore the dynamic aspect of the data by using a color scale which encode the time. Other layouts taking into account the time could be implemented in the future. The visual query system could also be extended by introducing more complex time constraints, such as in [**monroe2012exploring**].

5.8 . Conclusion and Future Work

We presented , a system for exploring temporal social networks aimed at social scientists. It relies on modeling data as bipartite, multivariate, dynamic social networks where persons are linked to documents or events using a typed link that expresses a role. We have successfully applied our data model to a wide variety of historical, sociological, genealogical datasets, and publication data. Our tool relies on this data model to allow historians to explore their data and then answer their sociological questions using 1) dynamic queries on the network structure to highlight groups of interests, and 2) visual comparisons to contrast selected groups according to their structure, time, or any other attribute. The results can be visualized as a node-link diagram, a geographical map, graph measures, and distributions of values for the attributes.

We have shown that complex explorations and analyses were easy to perform, and validated our approach by describing two use cases among many more projects

we are collaborating with.

By specifying a unifying data model and novel high-level visual and interactive tools for the comparison of topology, attributes, and time, we believe leads the way towards a new generation of highly interactive exploration tools applicable to analyze a wide variety of real social networks.

For future work, we need to improve the usability of , now that we know that the system is effective at exploring and analyzing real social networks. Our collaborators are eager to use it on their own and asked us to teach it to their students.

will be extended to support more SNA measures and computations such as clustering ; it would create a new attribute containing a cluster identifier. However, we are more interested in providing more layout options for the graph to better highlight the time, such as the PAOvis technique [**valdivia:hal-02264960**].