université PARIS-SACLAY

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

Analyse Visuelle pour l'Analyse de Réseaux Sociaux Historiques

Visual Analytics for Historical Network Research

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

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

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

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

Alexis PISTER

Composition du jury

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

THESE DE DOCTORAT

NNT : 2020UPASA001

université PARIS-SACLAY

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

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

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

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

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

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

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

Table des matières

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

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 [?]. 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 [?, ?, ?].

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 [?].

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

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 [?] and its language *Cypher*. Most visualization systems offer dynamic queries to hide the complexity of query languages. However, using a rich data model, some queries are much easier to refine using scripting than dynamic queries. We implemented dynamic queries that also show the translated Cypher queries, and inversely, can translate textual queries into visual queries. With that interface, social scientists can start building their queries with simple widgets and, if needed, complement them by editing the query, alone or with the help of power users. On top of that, they can easily copy and paste the textual query to share the current state of the query and associated results with someone else or to start an analysis session from a previous result. ComBiNet also implements subgraph comparison techniques, allowing the comparison of networks, network-related measures, and attribute distributions between the entities returned by the queries. We validate the query and comparison system with a usability study and we demonstrate ComBiNet can be used to answer sociological questions by describing in depth several real-world use cases.

After the related work section, we describe our data model in detail using four use cases, and present our system ComBiNet, with the design of the visual query and comparison features. Finally, we present 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.

1.1 . Related Work

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

1.1.1 . Graphlet Analysis

One of the inspiration of this project came after participating in the 2020 VAST challenge [1] where we used graphlets to measure similarity between several networks [?].

Graphlets are small connected induced, non-isomorphic subgraphs composing any network. In an induced subgraph, two vertices linked in the original graph remain linked in the subgraph. For instance, if the original graph is a triangle we can only induce the simple edge or triangle subgraph (graphlet) The path of length 2 has all vertices of the original graph but misses an edge and is, therefore, not a possible graphlet. They were first introduced by Milo et al. [?] to explore the structural differences between biological networks, but they are now used in several disciplines involving networks such as sociology.

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

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

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

1.1.2 . Visual Graph Querying

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

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

1.1.3 . Visual Graph Comparison

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

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

1.1.4 . Provenance

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

1.2 . Task Analysis and Design Process

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

1.2.1 . Use Cases

We elaborated this interface from the collaborations with historians we described in **??**. These collaborations involved regular meetings and multiple interviews over two years. All these datasets are textual corpora constituted of historical documents mentioning people with complex relationships. They are well modeled by bipartite multivariate dynamic network. We give more details about the datasets of these collaborations in this section and we also list our collaborators' main questions and the graph queries extracting the information to start answering them. The full answers involve visualizations of the query results and attribute summaries that we describe in the next section. We list the most important questions our collaborators shared with us on their respective datasets. We categorized those according to four dimensions : global (G)/local (L) (do they want to categorize group of nodes or retrieve specific persons/documents), if the question can be answered using the topology (T), and/or the attributes (A), and finally if a comparison (C) using several filters is needed or not (N).

check les questions

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

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

Table 1.1 – Tasks to support during exploration, according to our expert collaborators, split into 3 main high level tasks.

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

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

Query 1.1 Request all approvers occurrences

Query 1.2 Request all associates and guarantors occurrences

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

   Query 9.1 Select persons with a geolocated birth and death certificate

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

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

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

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

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

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

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

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

3. Analysis of migrations from Spain to Argentina through the **marriage acts at Buenos Aires in the 17–19th centuries (1396 acts, 6731 persons) [52].** The corpus is made of acts that mention the spouses and the witnesses of the wedding, which are the roles modeled by the links. The origin, date of birth, and parents' names are specified for both spouses.

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

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

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

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

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

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

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

4. Socio-political analysis of **migration of ethnic Germans from communist Romania to West Germany in the 20th century (ongoing work) [14].** The corpus is made of administrative forms that mention persons requesting to migrate, along with the persons they want to join, and the administrative persons of the ministry in charge of the forms (3 roles). The family members of the aspiring migrant are also mentioned in the forms, with their respective dates of birth.

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

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

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

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

1.2.2 . Tasks Analysis

Most of the questions we collected from our collaborators could be answered by isolating a subgroup of entities and analyzing them in the context of the whole network, or by comparing two subgraphs, in terms of their entities, structure, and attribute distributions. From discussions with our collaborators and the analysis of their questions on their data, we elaborated a list of requirements for the visual interface, split into three main parts : 1) Exploration of the data, 2) Queries, and 3) Comparisons. The elaboration of the tasks was an iterative process, as we showed the interface to our collaborators several times in the development phase to get feedback. The tasks are described here and summarized in Table 1.1 :

1. **Exploration of bipartite multivariate dynamic network**. The visual interface must allow exploration of this specific type of network, using every aspect of the data, i.e. its topology (T1.1), node attributes (T1.2), roles (T1.3), geolocation of the documents/events (T1.4) and time (T1.5). Common interactions such as selection and zooming are also needed for the exploration.
2. **Applying filters**. To answer their questions, users need to be able to apply filters to the data, to isolate specific groups of entities having specific behaviors or characteristics. To answer the diversity of questions, they should be

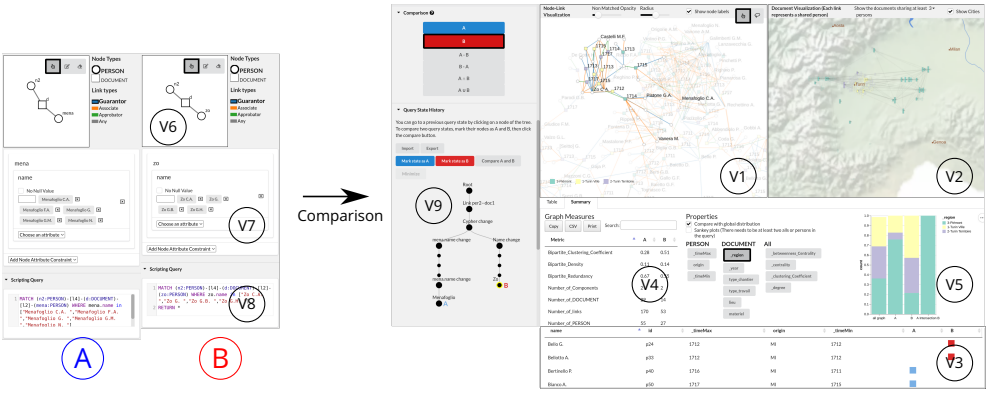


Figure 1.1 – The ComBiNet system used to compare two subgroups of a social network of contracts from [**?**], extracted with dynamic visual queries. (A) and (B) show the two visual queries created by the user in the query panel using an interactive node-link diagram editor (V6), dynamic query widgets (V7), and the equivalent Cypher script (V8). The right part shows ComBiNet's global interface in *comparison* mode : (V1) Network visualization panel, (V2) Map of the geolocalized nodes, (V3) Table of persons, (V4) Graph measures comparison, (V5) Attribute distribution plots, and (V9) Provenance tree. The two visual queries on the left, translated into Cypher queries below, select the "Menafoglio" family on the left, and of the "Zo" family on the right, along with their construction contracts and close collaborators.

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).

1.3 . The ComBiNet System

ComBiNet is designed to visualize, explore, and analyze social networks encoded as bipartite multivariate dynamic network. When started, it dynamically collects the node types, roles, sub-types, and attributes when reading the network from

the database. ComBiNet is constituted of four main panels, split in different views as shown in Figure 1.1 : the query and comparison panel, the graph visualization panel, the map visualization panel and the query results panel.

1.3.1 . Visualizations

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

V1 : Bipartite Node-Link Diagram The bipartite node-link visualization panel shows the network using the DrL force layout from igraph [?] with overlap removal using D3 [?]. Node-link representations are very common in social sciences [?,4,54] and were a specific request from our collaborators. In the context of our bipartite model, the persons are represented as circles and the documents/events as squares, while the roles are encoded as link colors. A link models the mention of a person in a document. This view provides an overview of the data by showing the structure of the network (T1.1) and the roles of the persons in their different documents (T1.2). Attribute values can be overlayed on the nodes using colors when users select an attribute. It allows detecting patterns relative to attributes, in the context of the topology of the network (T1.2, T1.4, T1.5). For example, Figure 1.2 shows the construction dataset of #1where the user selected the *year* attribute, coloring the documents nodes with their year in the node-link diagram (left). The view also provides pan & zoom and selection interactions for effective navigation.

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

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

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

V3 : Entities Tables All the persons and the documents of the loaded dataset are listed in two separate tables, showing the attributes of the entities. This way, users can order the entities according to any attribute they want (T1.2). The tables are

linked to the visualizations, meaning that selecting a row highlights the respective entity in the visualizations, and vice-versa. Tables in social network visualization systems have been proven to be efficient and useful for social scientists when exploring their data [?]. It allows them to link the visualization to the network entities more easily, and dive deeper on one entity's attribute values after selecting it in the network. It also makes ranking entities according various criteria easier and straightforward.

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

V5 : Attributes View All the attributes in the network are shown as buttons in the bottom right of the interface, sorted by their associated node type (person, document, and both). They can be quickly visualized by hovering over the button, producing two effects : it colors all the nodes on the two views according to their attribute values, and it shows a plot of the distribution of the selected attribute, as shown in Figure 1.2. By clicking on the button, the visual encoding and distribution remain selected. This interaction is inspired by the x-ray technique of the Vizster system [?]. Users can follow a first exploration of their data by visually detecting correlations between attribute values and some groups of persons or between attribute values and some specific areas in the map view (T1.2, T1.4, T1.5).

1.3.2 . Query Panel

The query panel allows to rapidly build queries visually, with topological and attribute constraints. The visualization of the query is synchronized with the Cypher query sent to the database. Modifying one representation will update the other, allowing users to build a query visually and refine it in Cypher when appropriate. Experts users who know the Cypher language can also start to construct their query textually and modify it visually later on. In this section, we describe all the features and interactions allowing ComBiNet to build a query and illustrate them with questions 2 and 6 of the use case #1. Our collaborator wants to *find the persons who are mutually Guarantor to each other in separate contracts (6)* and to know *how Torino and Torino's surroundings differ according to their contracts ?*.

V6 : Node-Link Dynamic Query

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

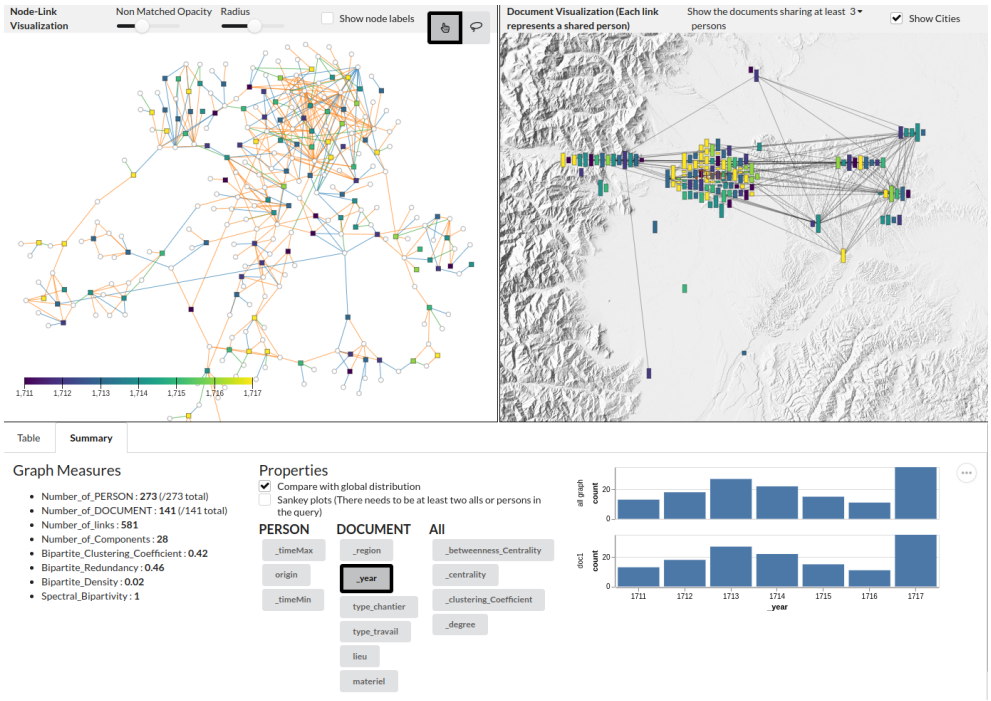


Figure 1.2 – ComBiNet interface with dataset of collaboration #1. The user selected the year attribute, showing the distribution of document years with an histogram (bottom right), and coloring the documents node on the bipartite view (left) and map view simultaneously (right).

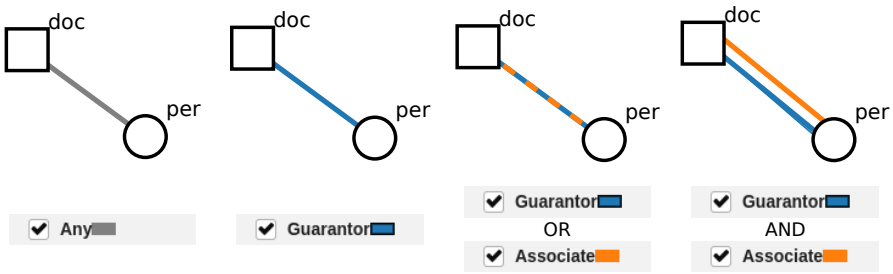


Figure 1.3 – All link creation possibilities : Any link type (left), one selected link type, here guarantor (middle left), union of several link types (middle right), several links with different types (right)

the nodes in the panel, while the *addition* and *deletion* modes allow the following actions :

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

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

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

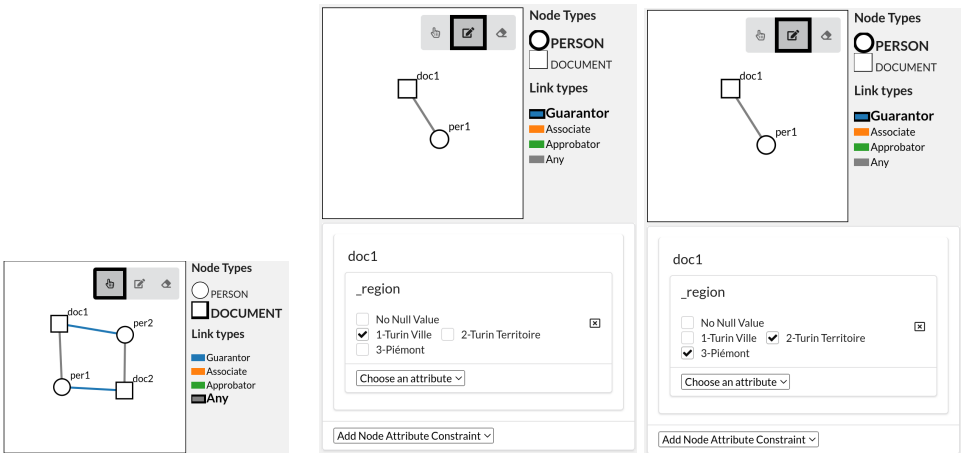
Link Creation : In *addition* mode, clicking on a node and dragging the mouse to another node will connect the two with a link. Its type (color) will be the link type selected on the legend.

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

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

Users build concrete subgraphs with the same representation as in the bipartite graph view : a visual query is a graph template. Each role (link type) is rendered using a color (Figure 1.3 left). We can also create untyped links using the *Any* value, which will be matched by all the existing link types (Figure 1.3 left). We also allow creating links that can be matched by several selected link types in the graph, by checking several possible types for one link. These links are represented by a dashed line with the colors of the possible types (Figure 1.3 middle right). Several links with different types can also be created among two nodes to query a person with more than one role in the same event (Figure 1.3 right). When a node or link is created in the query, it is given an identifier starting with *per* for a person, *doc* for a document, *link* for a link, followed by a number. These identifiers are used in the attribute constraint panels and the textual query and can be changed through their textual representations.

To find persons who are mutually guarantors in our collaboration #1, we first create one person and two documents using the addition mode and by clicking on



(a) (b)

Figure 1.4 – Visual queries created to answer questions 2 and 6 of our collaboration #1. (a) The visual query retrieve individuals who are mutually guarantor to each other in separate construction contracts. (b) The two visual queries retrieve the documents—along with the signatories—of *Turin* (left) and of *Turin Territoire* (right)

the canvas. We then link the person node to the first document with a link that is not typed (Figure 1.3 left), and link it to the second document with a Guarantor link (Figure 1.3 middle left). We then create a second person node and link it to the two documents with opposite link types. The resulting visual query is presented in Figure 1.4 (a). To answer the second question, we can simply start to request all the links in the graph, no matter the type, as shown in Figure 1.4 (b). The database will then return all the links in the graph with their attached nodes.

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

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

For the categorical and nominal widgets, selecting several values will correspond to the union of the filters. The three widget types are shown in Figure 1.5.

To answer our collaborator's second question (*how Torino and Torino's surroundings differ according to their contracts ?*), we first want to filter the documents which are located in Torino (*Turin* in French). For this, we start by selecting the whole dataset by linking a person and document node with *any* link. Then, we select the id *doc1* of the document of our visual node-link query, and the *region* attribute. It will initialize a categorical widget including all the values found in the dataset for this attribute with associated checkboxes. We check the region of interest "*1-Turin Ville*" to select all the documents from this region. The first widget of Figure 1.5 illustrates the created constraint. To select the documents of Torino's surroundings, we can simply uncheck the "*1-Turin Ville*" value for the *region* attribute and check the two other values "*2-Turin Territoire*" and "*3-Piemont*" which are areas corresponding to the surroundings of Torino. Both queries are represented in Figure 1.4 (b).

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

Query Results Each modification of the query, whether from the node-link dynamic query, the widgets, or the Cypher text boxes, update the two visualization panels (V1, V2), the entities tables (V3), the graph measures view (V5), and the attribute plots (V6). The nodes and links that do not match (are not retrieved by the query) are grayed out in V1 and V2, and are removed from the persons and documents tables (V3). A third table shows every occurrence found of the created pattern that we call the occurrence table. The occurrence table for question 1 of collaboration #1is shown in Figure 1.6 (a). It tells us that the pattern has been found 36 times. Users can switch between the three tables in the table view using the tabs. The graph measures are computed on the new graph formed by the union of all patterns found and updated on the graph measures view (V5). Figure 1.6 (b) (left) shows the user the different graph measures of the subgraph induced by the patterns found. Since some measures can be long to compute, the values are computed iteratively in the backend and shown progressively [**?**] to avoid blocking the interface. The distribution plots in the attributes view (V6) are updated, showing the values of the entities of the latest constructed query, next to the global distributions.

Attributes Visualization When users select an attribute in the attributes view (V5), its distribution is visualized for the queried entities and the whole network. However, these plots show the aggregated values and we lose the potential value transitions between the query nodes. For example, Figure 1.7 shows a query to list the persons with the role of "approbator" (green) in a contract after being a

"guarantor" (blue) in another contract (using a time constraint). We may want to see if the locations or types of the two contracts are the same or if they change, case by case. Unfortunately, we lose this information with the aggregated plots. By checking the "Sankey" option on top of the distribution visualization, the plots are transformed into Sankey diagrams, giving information on how the attribute values relate between the nodes (person or event) of the same query. A Sankey diagram showing the attribute distributions is particularly useful for queries where the nodes have intrinsic time relationships, such as birth certificates, marriage, or death certificates where we know the order in which these events occurred. It is also useful for queries with user-defined time order constraints as in Figure 1.7. The graph measures and attribute visualization view for the results of question 1 of collaboration #1are shown in Figure 1.6. The sankey view of the *origin* attribute show that mutual guarantors come from 4 regions only, and that usually people have mutual guarantor relationships only with persons of the same origin. This is especially true for persons from *Milano*, and with some reciprocal links between persons from *Bioglio* and the *Comune di Ro*.

V9 : Provenance Tree Each change in the query panel is saved with the computed results so that the history of the query construction can be shown in the form of a provenance tree (T2.4), managed using the Trrack library [**?**]. Each node of the tree represents a query change, with a description label like "New Link". It allows to rapidly visualize the succession of filters applied with their refinements. At any moment, users can click on a tree node to go back to the previous state ; allowing to navigate in the exploration states. Hovering over a node shows a tooltip with the query panel associated with the selected query state. It let users rapidly see what query is associated with each node of the tree If a new change is made on the query from a previous state, a new branch is created on the tree, allowing to revisit and refine explorations. Figure 1.8 shows the provenance tree made to answer question 2, split in 2 branches, with the tooltip showing one of the node query state.

1.3.3 . Comparison

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

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

For the second example from Table 1.1, we want to compare the works in Torino with the ones in Torino surrounding. Since we previously constructed the query returning all the contracts from *Turin* with the mentioned people, we can

return to this point in the provenance tree, and change the constraint of the *region* attribute from *1-Turin Ville* to *2-Turin Territoire* and *3-Piemont* using the checkbox to get the two queries we want to compare. They are shown in Figure 1.4. The user can then rename the provenance tree nodes with explicit names such as "Torino" and "Surroundings", and mark them as A and B using the appropriate buttons. Clicking on the "Compare State A and B" will make the interface compare the two query results.

Topological Comparison In visualization mode, users can rapidly switch between the visual filters of (A) and (B) by hovering over their respective buttons on the comparison menu and thus compare the structure of the two resulting subgraphs (T3.1). Similarly, different boolean comparison operations are available by hovering their respective buttons (Figure 1.1-C), such as the intersection, union, and differences between the two filters. Moreover, the summary tab (top of Figure 1.1-D) allows comparing the different graph measures of the two subgraphs by showing them side by side (T3.3). Comparing these measures, such as the number of matched documents or the densities, is crucial for SNA.

Check la comparaison

comparaison de torino et territoire sur le graphe

Attribute-Based Comparison The comparison of one or several attribute distributions between (A) and (B) is also useful for answering the historical questions of our users. In the attribute view (V5) of the results panel, hovering or clicking on an attribute name will show the distribution of this attribute in four contexts : the nodes of the whole graph, the queries (A), (B), and the currently selected Boolean operator (e.g., intersection or union) if one is selected. This allows users to compare attribute distributions between several subsets of interest (T3.2). For example, we can compare the attributes between the contracts of Torino and the ones of its surroundings. We can also compare the persons who worked in Torino, in Torino's close territory, and in both areas, by selecting the intersection operator. Figure 1.10 illustrates the comparison charts for different attributes. We can see that the types of construction sites differ between the two regions : the city of Torino clearly has a lot of military sites compared to the surroundings of Torino, which has almost none. This is the opposite for the number of religious sites, which are almost all localized in the surroundings of Torino. If we now look at the year distribution of the contracts, we can see a difference in the the distributions. The years of Torinos's construction contracts were steady between 1711 and 1717 with a little spike in 1713, while the constructions were more scare in the surroundings before 1716. We can see a big spike of constructions in 1717. This is interesting to our users, as it shows the dynamic of the construction in the area : the center of the city started to be constructed before other constructions arisen in the surroundings.

We can also compare the profile of persons who collaborated at Torino and Torino surroundings by selecting the intersection of those two queries. One of the

questions the historian had (question 2 of Table 1.1) was to know if those persons were a group with specific attributes and characteristics, or were inseparable from other persons working in the two areas. If we look at the betweenness centrality, on average, the values are higher for this group of people, meaning that the persons who work on the construction site at Torino and Torino's territory are clearly two distinct groups, and the persons collaborating in the two areas act as bridges between these groups. This visual demonstration was convincing and revealing for our users.

1.3.4 . Implementation

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

1.4 . Use Cases

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

1.4.1 . Construction sites in Piedmont (#1)

One of the main questions of our collaborator was to compare two families which he knew played a big role in the structure of the network : the *Menafoglio* and *Zo* families (question 4 in Table 1.1). Specifically, he was interested in knowing if there were differences in specialization in type of contracts and area of work for the core members of these families, and to what extent the two families were collaborating. Moreover, he was very interested in characterizing the group of people collaborating with both families.

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

We then do the same query for the Zo family. We keep the same topological filter and replace the name filters with the core members of the Zo family known by the historian. We see on the graph view (Figure 2 of the supplementary material)

that the group is smaller and is on a different area in the graph. The map enriched with a selection of the *region* attribute shows that, contrary to the Menafoglio, the Zo have been more active in Turin and around.

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

Our user is then interested in comparing the distribution of several attributes between the two groups. We can clearly see in Figure 1.11 (middle) 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.

1.4.2 . French Genealogy (#2)

We describe how ComBiNet allowed to answer an important questions of the use case #2 : to detect the largest migrations across several generations, in which areas, and at what time they occured (question 7 in Table 1.1). The map view shows at a glance (Figure 3 in the supplementary material) that the majority of events has taken place in three specific regions west, mid-north, and mid-south.

To find patterns of migrations within families, we first make a query representing a simple family by linking a person node to a birth event, connected to the parents using a link of *father* or *mother* type. We repeat the process to the new parent node to add another generation. Finally, we connect the latest generation child with a death event, to have another date and location to compare to (see Figure 1.12a). This query returns every person with their parents and grandparents, along with their respective birth and death data for the latest person. We also create a constraint on the *department* attribute on the documents to only retrieve the events that have a non-null associated location. This request returns a subgraph of 64 persons and 88 documents. The user can now select the *department* attribute to create a Sankey diagram that shows the change of departments across the different generations of the families. Figure 1.12b shows that the majority of families are from *Haute-Vienne* (which can easily be confirmed by checking the map), and do not move much across generations. Our collaborator however detected interesting patterns of people moving from the department *Creuse* to *Haute-Vienne* across two generations. She refined the query by adding an attribute filter on this specific department using a widget. The table view then showed her who these migrants were and when it occurred. The bipartite visualization panel allowed exploring more in-depth this specific group of people.

Afterward, we answered the question 8 (Table 1.1), to compare the migrations between the 18th and 19th centuries. She thought people started moving in the 19th century and wanted to confirm it. To answer this, we first created a query to retrieve the people with birth and death certificates from a specified department. We then applied a time filter on the death certificate node, first for the 18th century and then the 19th century, compared the two query results using the comparison mode, and looked side by side the Sankey graphs related to *departments* (Figure 1.13). We can clearly see that people do not move at all in the 18th century, while in the 19th century even if the majority of people stay in the same place from their birth to their death, more than half move.

1.4.3 . Sociology thesis in France

We describe in this third use case how ComBiNet can be used to answer questions about thesis in France between 2016 and 2022. Indeed, some sociological datasets made of documents can also be well modeled as bipartite multivariate dynamic networks like for example thesis dissertations : a thesis is a document with specific attributes such as the subject, the doctoral school, the domain, the university, and the date of defense, and mention several peoples who are socially connected through the thesis defense with different roles : author (*auteur* in french), director(s) (*directeur*), referees (*rapporteur*), and jury president (*président de jury*). We present here an exploration of the data by ourselves using ComBiNet. A first look of the graph measures tells us that 896 thesis have been defended in sociology in France between 2016 and 2021 in France, with 2453 persons included in the defenses (see Figure 1.14 bottom). The bipartite node-link view shows us an overview of the network, but is hard to parse due to the network's size. Zoom actions though allow to center the view for specific parts of the network. The map view allows us to see that thesis have been defended all around France. We can however see that the majority of thesis are defended in Paris. This is confirmed if we look at the distribution of the cities (Figure 1.14 bottom right) : around half of the defenses are in Paris, compared to the rest of the country which is more or less homogeneous. By setting the threshold to link creation to one (meaning that a link is created between two documents if they mention at least one common person), a lot of links are created as seen in Figure 1.14 (right). It means that a lot of thesis defenses include referee and juries from different cities.

Let's now try to answer an interesting question : "Do referees and jury presidents often ask thesis directors to be referees and jury presidents in their turn of other thesis where they are directors ?"". For this, we can construct a visual query representing this pattern by creating two person nodes and two document nodes, and by connecting them with two president links and two referee or jury director links in a symmetrical way, as shown in Figure 1.15 (right). The occurrence table tells us that this pattern has been found 76 times in the network, meaning that this is a recurrent behaviour. We are now interested in characterizing the thesis occurring in this pattern, by their regions. We can look at the *city* attribute distribution

for this thesis by selecting it in the attribute view as shown in Figure 1.15 (bottom right). We can first see on the map that this pattern occur mainly in the biggest cities of the country. By selecting the Sankey view option, we can investigate if this pattern occur between thesis defended in different regions or if it occur mainly in the same ones. We learn that it depends mainly of the regions : in Bourgogne-Fanche-Comté 26 out of 29 thesis are connected with thesis of another region. In contrary in *Occitanie* it is the case for only 4 out of 17. In average, we can see that this pattern occur a lot for thesis of the same region. In Ile-de-France, it is the case for around half of the thesis (28/50). This explorative analysis shows that ComBiNet can be use to explore and gain insight on such datasets.

1.5 . Formative Usability Study

We performed a formative usability study with two historians and one expert in visualization. We had 3 meetings with each, and gave them control of the tool to see if they could use it to explore their data, 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 redesign of some core features, like the activation of the comparison mode which is now started by first marking the state nodes in the provenance tree. It also led to the implementation of new features, such as the person and document tables (which are updated after each query), the persistent selection of nodes across the two views and the tables, and the undo feature for visual queries. At the final meetings, the three users were able to perform exploration, queries and comparisons to answer socio-historical questions by themselves.

1.5.1 . Feedback

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

1.6 . Discussion

Query Expressiveness. The visual query system currently allows finding occurrences of attributed subgraphs, with potential union operations on constraints (links

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

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

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

1.7 . Conclusion and Future Work

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

describing two use cases among many more projects we are collaborating with and by performing a formative usability study showing that the system is usable by social scientists.

By specifying a unifying data model and novel high-level visual and interactive tools for comparing topology, attributes, and time, social scientists were able to clean their data more easily by finding errors and inconsistencies by exploring the network and querying errors induced patterns. Thanks to the document-centered model, it was easy for them to trace back the errors and inconsistencies to the sources for corrections. With the same representation they were able to operate explorations and analyses using complex interactions implemented in ComBiNet such as coordinated views, visual querying and comparisons mechanisms.

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

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

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

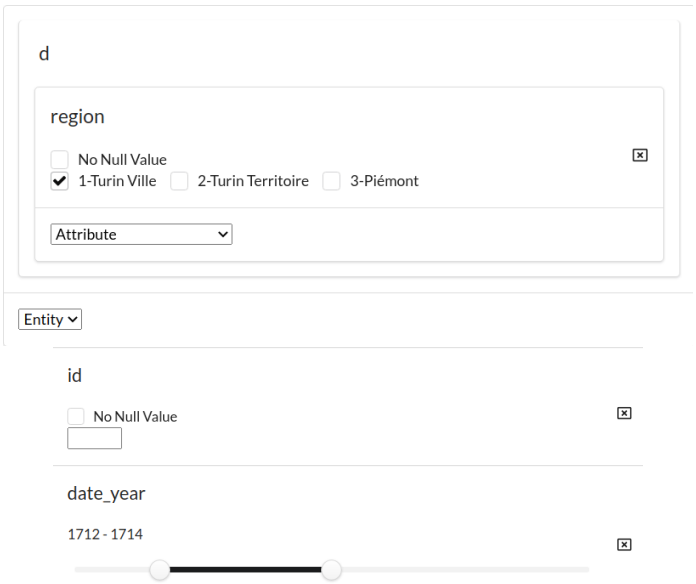
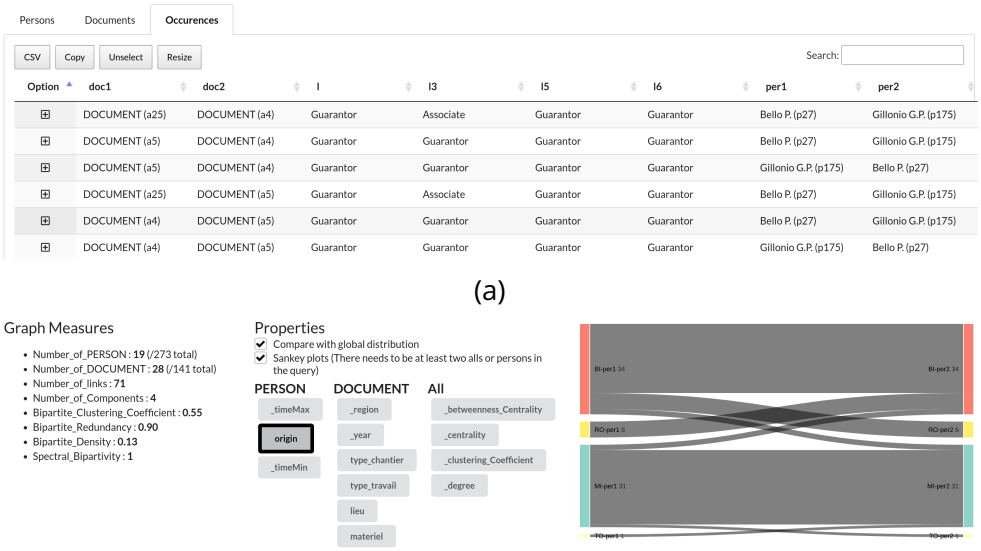


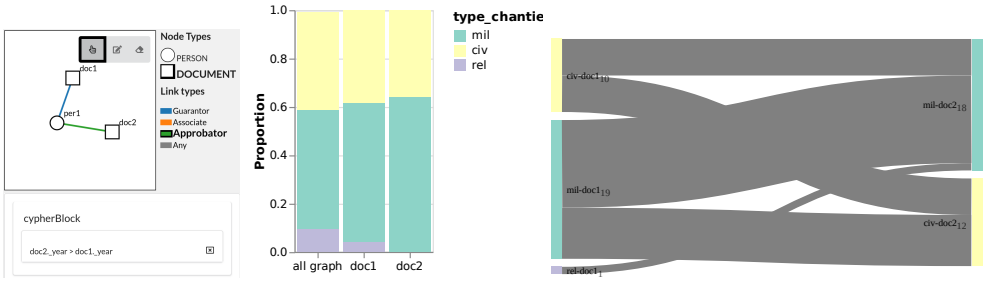
Figure 1.5 – Widget designs for the different attribute types : checkboxes for categorical attributes (top), text input for nominal attributes (middle), and double slider for numerical attributes (bottom). The categorical attribute example shows the options input letting users create new constraints for other attributes and other nodes.



(a)

(b)

Figure 1.6 – Results of the question 1 of collaboration #1 : (a) shows a subset of the table view with every occurrence of the pattern found. (b) shows the summary panel, with the graph measures and the attributes view with the *origin* attribute selected and the sankey option checked. It allows to see the attribute distribution of the persons included in the pattern, and see if there is a relationship between persons who are mutually guarantors and their origin.



(a) (b) (c)

Figure 1.7 – Two ways of showing the distribution of "type chantier" (type of works), a categorical attribute with three possible values "*reli*gious", "*mil*itary", and "*civ*ilian". (a) A query matching the contracts made by the same person (*per1*) as an "approbator" (green link to *doc2*) after being a "guarantor" (blue link to *doc1*) using the constraint (`doc2._year > doc1._year`). (b) Stacked bar chart for the matches, the earlier contract (*doc1*), and the older contract (*doc2*), and (c) Sankey diagram with the early values on the left and the last on the right. The Sankey diagram reveals the value changes between the two documents : the guarantor who worked initially on religious work switched to military work.

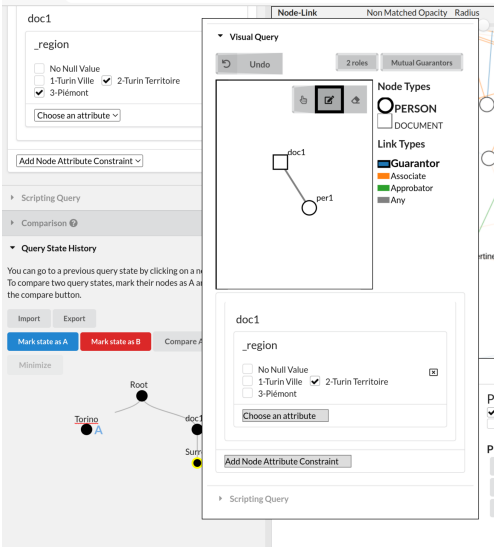


Figure 1.8 – Provenance tree to answer question 2 of collaboration #1 : left branch lead to Torino documents (the node is labeled as A) while right branch lead to surrounding documents (the node is labeled as B). The user hover over one node, revealing a tooltip which shows the visualization of the node's query..

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

Figure 1.9 – Comparison table of the graph measures the query filters (A) and (B)

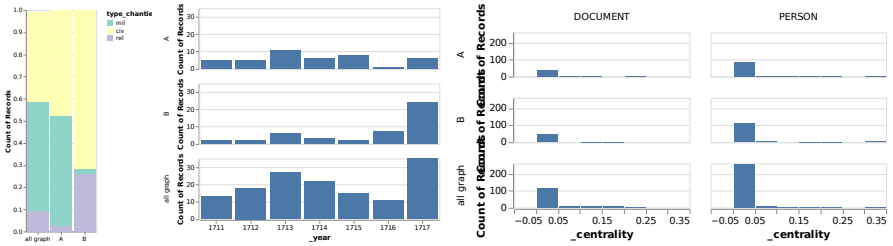


Figure 1.10 – Distribution of the type of constructions, the years and the centrality for the documents and signatories of Torino (A), Torino surroundings (B), and the whole graph. (top).

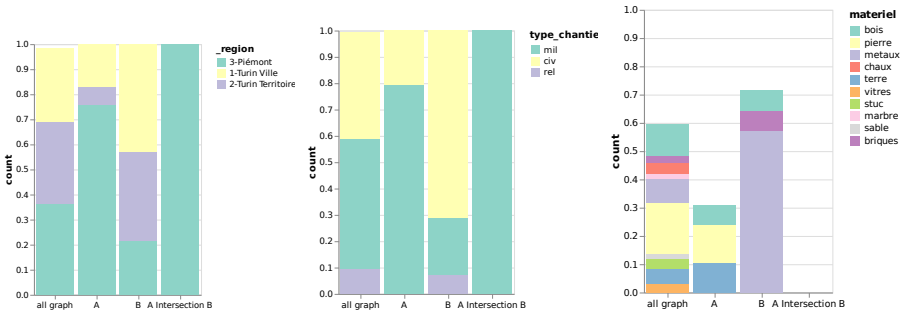
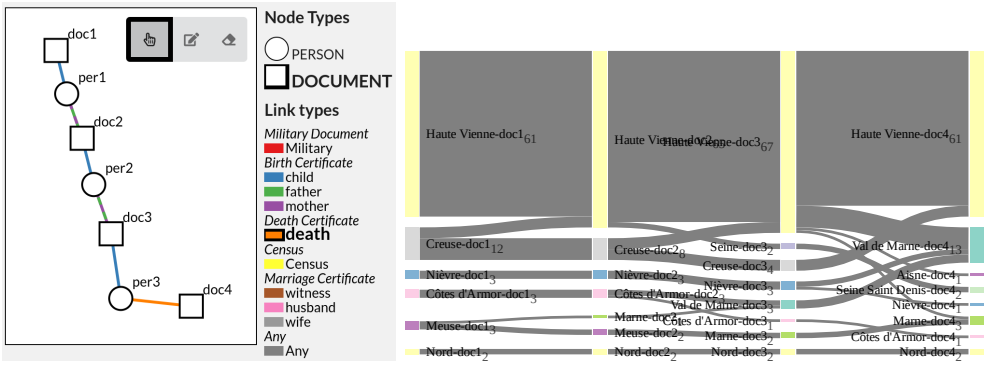


Figure 1.11 – Attributes distributions plots between the whole graph, the *Menafoglio* family (A), the *Zo* family (B), and $A \cap B$, for the *region,type_chantier, materiel type.*



(a) Visual query to find all 3-generation families

(b) Sankey diagram showing the birth and death places of people across generations

Figure 1.12 – Migrations across departments over three generations

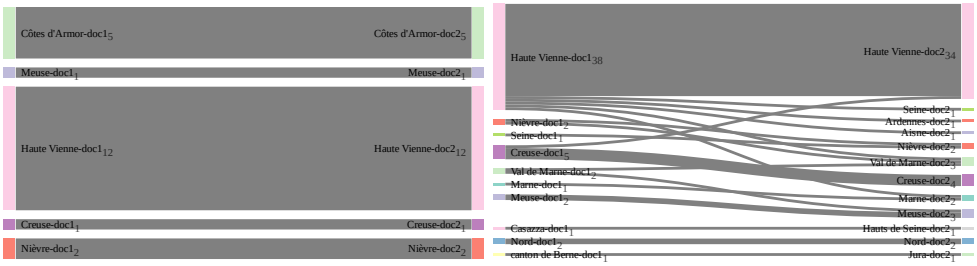


Figure 1.13 – Sankey diagrams showing the migration of people in the 18th and 19th centuries, extracted from their birth and death places.

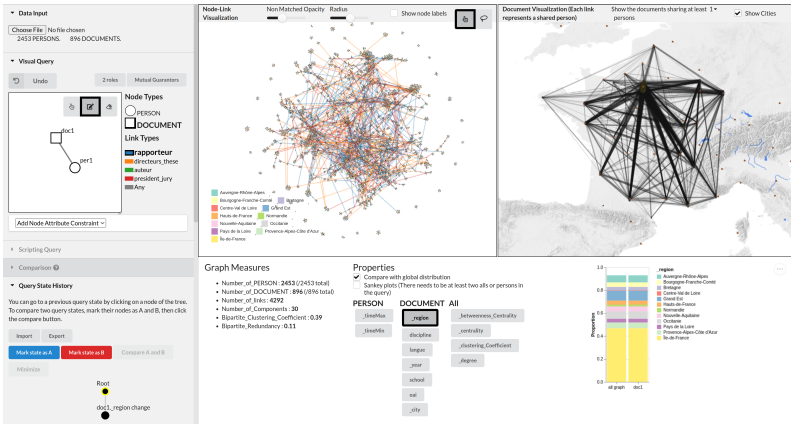


Figure 1.14 – ComBiNet used for exploring thesis of sociology defended in France between 2016 and 2021. The bipartite and map views show an overview of two vision of the network. The user select the *region* attribute, showing the geographical distribution of the defended thesis.

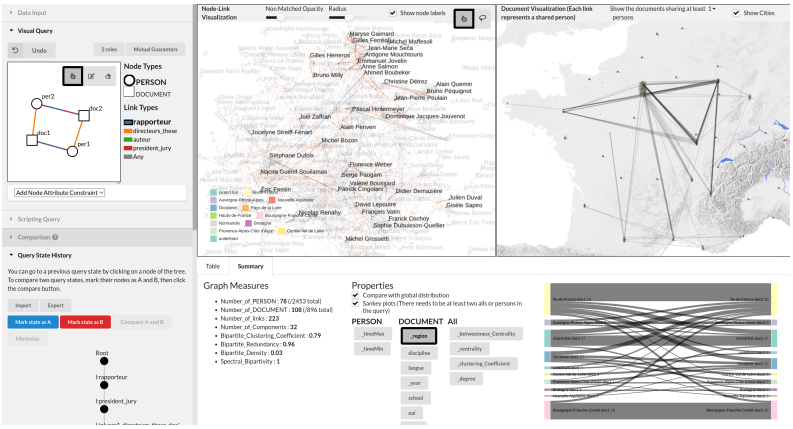


Figure 1.15 – Sociology thesis dataset explored with ComBiNet. The user constructed a visual query to see there are symmetrical relationships between thesis directors and referees (or jury directors). The *region* attribute is selected with the Sankey option, letting the user see if their are correlations between the regions of the thesis found in this pattern.

Bibliographie

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

[2] Mashael AlKadi, Vanessa Serrano, James Scott-Brown, Catherine Plaisant, Jean-Daniel Fekete, Uta Hinrichs, and Benjamin Bach. Understanding barriers to network exploration with visualization : A report from the trenches. *IEEE Trans. Vis. Comput. Graphics*, 27(2), February 2023.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[23] GEDCOM : The genealogy data standard.

[24] Mohammad Ghoniem, J.-D. Fekete, and Philippe Castagliola. A comparison of the readability of graphs using node-link and matrix-based representations. In *IEEE Symposium on Information Visualization*, pages 17–24. Ieee, 2004.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[66] Marc A. Smith, Ben Shneiderman, Natasa Milic-Frayling, Eduarda Mendes Rodrigues, Vladimir Barash, Cody Dunne, Tony Capone, Adam Perer, and Eric Gleave. Analyzing (social media) networks with NodeXL. In *Proceedings of the Fourth International Conference on Communities and Technologies*, C&T '09, pages 255–264, New York, NY, USA, June 2009. Association for Computing Machinery.

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

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

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

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

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

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