======= I conclude with discussions on the potential merging of both tools, and more globally on research directions towards better integration of visual analytics systems on the whole workflow of social historians. Systems with a focus on those properties—traceability, simplicity, and document reality—can limit the introduction of bias while lowering the requirements for the use of quantitative methods for historians and social scientists which has always been a controversial discussion among practitioners. »»»> 61c6b43c8379c08ed6640dc5f5fef5a79dd20b80

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

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

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

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

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

Thèse soutenue à Paris-Saclay, le 15 décembre 2022, par

Alexis PISTER

Composition du jury

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

THESE DE DOCTORAT

NNT : 2020UPASA001

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

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

Mots clés: analyse visuelle, analyse de réseau sociaux, visualisation de réseaux sociaux, histoire sociale, réseaux historiques

Résumé: Cette thèse vise à identifier théoriquement et concrètement comment l'analyse visuelle peut aider les historiens dans leur processus d'analyse de réseaux sociaux. L'analyse de réseaux sociaux est une méthode utilisée en histoire sociale qui vise à étudier les relations sociales au sein de groupes d'acteurs (familles, institutions, entreprises, etc.) en reconstruisant les relations du passé à partir de documents historiques, tels que des actes de mariages, des actes de naissances, ou des recensements. L'utilisation de méthodes visuelles et analytiques leurs permet d'explorer la structure sociale formant ces groupes et de relier des mesures structurelles à des hypothèses sociologiques et des comportements individuels. Cependant, l'inspection, l'encodage et la modélisation des sources menant à un réseau finalisé donnent souvent lieu à des erreurs, des distorsions et des problèmes de traçabilité, et les systèmes de visualisation actuels présentent souvent des défauts d'utilisabilité et d'interprétabilité. En conséquence, les historiens ne sont pas toujours en mesure de faire des conclusions approfondies à partir de ces systèmes : beaucoup d'études se limitent à une description qualitative d'images de réseaux, surlignant la présence de motifs d'intérêts (cliques, îlots, ponts, etc.). Le but de cette thèse est donc de proposer des outils d'analyse visuelle adaptés aux historiens afin de leur permettre une meilleur intégration de leur processus global et des capacités d'analyse guidées. En collaboration avec des historiens, je formalise le processus d'une analyse de réseau historique, de l'acquisition des sources jusqu'à l'analyse finale, en posant comme critère que les outils utilisés dans ce processus devraient satisfaire des principes de traçabilité, de simplicité et de réalité documentaire (i.e., que les données présentées doivent être conformes aux sources) pour faciliter les va-et-vient entre les différentes étapes et la prise en main par l'utilisateur et ne pas distordre le contenu des sources. Pour satisfaire ces propriétés, je propose de modéliser les sources historiques en réseaux sociaux bipartis multivariés dynamiques avec rôles. Ce modèle intègre explicitement les documents historiques sous forme de nœuds, ce qui permet aux utilisateurs d'encoder, de corriger et d'analyser leurs données avec les mêmes outils. Je propose ensuite deux interfaces d'analyse visuelle permettant, avec une bonne utilisabilité et interpretabilité, de manipuler, d'explorer et d'analyser ce modèle de données. Le premier système ComBiNet offre une exploration visuelle de l'ensemble des dimensions du réseau à l'aide de vues coordonnées et d'un système de requêtes visuelles permettant d'isoler des individus ou des groupes et de comparer leurs structures topologiques et leurs propriétés. L'outil permet également de détecter les motifs inhabituels et ainsi de déceler les éventuelles erreurs dans les annotations. Le second système, PK-Clustering, est une proposition d'amélioration de l'utilisabilité et de l'efficacité des mécanismes de clustering dans les systèmes de visualisation de réseaux sociaux. L'interface permet de créer des regroupements pertinents à partir des connaissances a priori de l'utilisateur, du consensus algorithmique et de l'exploration du réseau dans un cadre d'initiative mixte. Les deux systèmes ont été conçus à partir des besoins et retours continus d'historiens, et visent à augmenter la traçabilité, la simplicité, et la réalité documentaire des sources dans le processus d'analyse de réseaux historiques. Je conclus sur la necessité d'une meilleure intégration des systèmes d'analyse visuelle dans le processus de recherche des historiens. Cette intégration nécessite des outils plaçant les utilisateurs au centre du processus avec un accent sur la flexibilité et l'utilisabilité, limitant ainsi l'introduction de biais et les barrières d'utilisation des méthodes quantitatives, qui subsitent en histoire.

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

Keywords: visual analytics, social network analysis, social network visualization, social history, historical networks

Abstract: This thesis aims at identifying theoretically and concretely how visual analytics can support historians in their social network analysis process. Historical social network analysis is a method to study social relationships between groups of actors (families, institutions, companies, etc.) through a reconstruction of relationships of the past from historical documents, such as marriage acts, migration forms, birth certificates, and censuses. The use of visualization and analytical methods lets social historians explore and describe the social structure shaping those groups while explaining sociological phenomena and individual behaviors through computed network measures. However, the inspection and encoding of the sources leading to a finalized network is intricate and often results in inconsistencies, errors, distortions, and traceability problems, and current visualization tools typically have usability and interpretability issues. For these reasons, social historians are not always able to make thorough historical conclusions: many studies consist of qualitative descriptions of network drawings highlighting the presence of motifs such as cliques, components, bridges, etc. The goal of this thesis is therefore to propose visual analytics tools integrated into the global social historians' workflow, with guided and easy-to-use analysis capabilities. From collaborations with historians, I formalize the workflow of historical network analysis starting at the acquisition of sources to the final visual analysis. By highlighting recurring pitfalls, I point out that tools supporting this process should satisfy traceability, simplicity, and document reality principles to ease bask and forth between the different steps, provide tools easy to manipulate, and not distort the content of sources with modifications and simplifications. To satisfy those properties, I propose to model historical sources into bipartite multivariate dynamic social networks with roles as they provide a good tradeoff of simplicity and expressiveness while modeling explicitly the documents, hence letting users encode, correct, and analyze their data with the same abstraction and tools. I then propose two interactive visual interfaces to manipulate, explore, and analyze this data model, with a focus on usability and interpretability. The first system ComBiNet allows an interactive exploration leveraging the structure, time, localization, and attributes of the data model with the help of coordinated views and a visual query system allowing users to isolate interesting groups and individuals, and compare their position, structures, and properties. It also lets them highlight erroneous and inconsistent annotations directly in the interface. The second system, PK-Clustering, is a concrete proposition to enhance the usability and effectiveness of clustering mechanisms in social network visual analytics systems. It consists in a mixed-initiative clustering interface that let social scientists create meaningful clusters with the help of their prior knowledge, algorithmic consensus, and interactive exploration of the network. Both systems have been designed with continuous feedback from social historians, and aim to increase the traceability, simplicity, and document reality of visual analytics supported historical social network research. «««< HEAD I conclude with discussions on the potential merging of both tools, and more globally on research directions towards a better integration of visual analytics systems on the whole workflow of social historians. Systems with a focus on those properties—traceability, simplicity, and document reality—can limit the introduction of bias while lowering the requirements for the use of quantitative methods for historians and social scientists which has always been a controversial discussion among practitioners.

Aknowledgments

Those three years went incredibly fast, and yet I feel I grew tremendously during that time both as a person and researcher, thanks to the many persons I encountered who supported, inspired, and motivated me.

I would first like to thank my two advisors Jean-Daniel and Christophe, for guiding me in those three years as great mentors. Jean-Daniel, for his continuous and generous support which tremendously helped me navigate those three years with excitement and joy, even during moments of doubt and unproductiveness. His door was always open to me for sharing my ideas and thoughts, which always led to great humane and scientific discussions and never lacked to motivate me. I will always be proud of having been his student. Christophe, for his continuous welcomed feedback. I started truly discovering the world of social sciences through him, and he always knew how to guide my work, especially when I was going in too complex and abstract directions. For this thank you.

I also thank Guy Melançon and Ulrik Brandes for accepting to review my manuscript on their time, and Wendy Mackay, Laurent Beauguitte, and Uta Hinrichs to have accepted being part of my jury.

This thesis would not exist without the collaborations, discussions, and exchanges I had with several social scientists and historians. Particularly, I want to thank Pascal Cristofoli and Nicole Dufournaud for their tremendous generosity, help, feedback, and support; this thesis would not have been possible without you. I would also like to thank Dana Diminescu and Zacarias Moutoukias for the insightful discussions and the data they shared with me. Finally, I would like to thank Claire Lemercier for her precious feedback and the useful references she pointed out to me.

My thoughts also go to all the people of the Aviz team with whom I shared my last three years. It was always a pleasure to come work at the lab in such a great atmosphere. I would first like to thank Natkamon Tovanich, for all the beers and guilty pleasures of complaining about research life, Gaëlle Richer, for all the great and insightful discussions, and Paola Valdivia, for her helpful guidance. It was fun to navigate (and get lost a couple of times) in the VAST challenge with you, and I feel I learned a lot by working on your side. I also want to thank Mickael Sereno, for his sharing of technical knowledge (even though you always go a little too fast on the bike), and Alaul Islam, for his good mood and kindness. It was a pleasure to be your office neighbor for three years, as discussing, laughing, and complaining with you about administrative issues always constituted agreeable work breaks. I also thank Jiayi who read a part of my manuscript (with Alaul and Natkamon), Sara for her inspirational productivity, and all the others who are or have been part of the team: Petra, Tobias, Lijie, Federica, Nivan, Sarkis, Marie, Xiayo. Finally, thanks to Catherine Plaisant, who always gave me great and appreciated feedback on my work, and Paolo Buono for his welcomed help on servers and deployments—two things I spent way more time on than I thought I would.

I also want to thank all of my friends from Paris, who allowed me to escape work life when needed and spend pleasing times. Kaelan, who always has truthful words, Paul, who always have kind words, Bastien for his mix of craziness and kindness (I should have more time to answer calls now), Pierre, whose complaining never fails to make me laughs, Aurélie, for her appreciated extroversion, Jurgen for the great discussions. Hiking, playing music, going on vacation, and partying with all of you never failed to make me happy, and feel grateful to have you in my life.

My thoughts also go to my parents, who conveyed to me their open-mindedness and curiosity.

Last but not least, I thank Julia, my love, who always helped and supported me in those three years, even in the most difficult times, and who even read a part of this manuscript. I could not dream of a better partner as you continuously bring joy and laughter into my life, and never failed to cheer me up and motivate me when I was feeling down. I am grateful to have you. After all those years I feel our bond is now stronger than ever, and I am looking forward to the next adventures that await us.

On the usage of footnotes

This thesis led me to read many history books and articles, in which footnotes are more than widespread. I grew rather fond of them, which explains why you will see several of them across this manuscript.

On the usage of the pronouns we and I

Most of the research described in this thesis was highly collaborative. I would like to thank deeply all my collaborators for their help, support, and thoughtful discussions. In the writing, I hence use "we" for collaborative parts and "I" for the parts I have mostly done myself.

Contents

1 Introduction **1**

1.1 Social History and Historical Social Network Analysis . . . 2
1.2 Visualization and Visual Analytics . . . 4
1.3 Visual Analytics Supported Historical Network Research . . . 5
1.4 Contributions and Research Statement . . . 9

2 Related Work **11**

2.1 Visualization . . . 12
2.1.1 Information Visualization . . . 12
2.1.2 Visual Analytics . . . 15
2.2 Quantitative Social History . . . 16
2.2.1 History, Social History, and Methodology . . . 16
2.2.2 Quantitative History . . . 18
2.2.3 Digital Humanities . . . 19
2.3 Historical Social Network Analysis . . . 20
2.3.1 Sociometry to SNA . . . 21
2.3.2 Methods and Measures . . . 23
2.3.3 Historical Social Network Analysis . . . 25
2.3.4 Network Modeling . . . 25
2.4 Social Network Visualization . . . 27
2.4.1 Graph Drawing . . . 28
2.4.2 Social Network Visual Analytics . . . 29

3 Historical Social Network Analysis Process, Pitfalls, and Network Modeling **33**

3.1 Context . . . 34
3.2 Related Work . . . 35
3.2.1 History Methodology . . . 35
3.2.2 Historian Workflows . . . 36
3.3 Historical Social Network Analysis Workflow . . . 37
3.3.1 Examples . . . 37
3.3.2 Workflow . . . 38
3.3.3 Visual Analytics Supported Historical Social Network Analysis . . . 40
3.4 Network Modeling and Analysis . . . 42
3.4.1 Network Models . . . 43
3.4.2 Bipartite Multivariate Dynamic Social Network . . . 46
3.5 Applications . . . 47
3.6 Discussion . . . 48
3.7 Conclusion . . . 48

4 **ComBiNet: Visual Query and Comparison of Bipartite Dynamic Multivariate Networks with Roles** **51**
4.1 Context 52
4.2 Related Work 54
4.2.1 Graphlet Analysis 54
4.2.2 Visual Graph Querying 55
4.2.3 Visual Graph Comparison 55
4.2.4 Provenance 56
4.3 Task Analysis and Design Process 56
4.3.1 Scenarios 56
4.3.2 Tasks Analysis 58
4.4 ComBiNet System 59
4.4.1 Visualizations 59
4.4.2 Query Panel 62
4.4.3 Comparison 69
4.4.4 Implementation 71
4.5 Use Cases 72
4.5.1 Construction Sites in Piedmont (#1) 72
4.5.2 French Genealogy (#2) 73
4.5.3 Marriage Acts in Buenos Aires (#3) 76
4.5.4 Sociology Theses in France 76
4.6 Formative Usability Study 78
4.6.1 Feedback 79
4.7 Discussion 80
4.8 Conclusion and Future Work 81

5 **PK-Clustering: Integrating Prior Knowledge in Mixed-Initiative Social Network Clustering** **83**
5.1 Context 84
5.2 Related Work 87
5.2.1 Graph Clustering 87
5.2.2 Semi-supervised Clustering 87
5.2.3 Mixed-Initiative Systems and Interactive Clustering 88
5.2.4 Groups in Network Visualization 88
5.2.5 Ensemble Clustering 88
5.2.6 Summary 89
5.3 PK-clustering 89
5.3.1 Overview 89
5.3.2 Specification of Prior Knowledge 91
5.3.3 Running the Clustering Algorithms 91
5.3.4 Matching Clustering Results and Prior Knowledge 92
5.3.5 Ranking the Algorithms 93
5.3.6 Reviewing the Ranked List of Algorithms 94

5.3.7 Reviewing and Consolidating Final Results 95
5.3.8 Wrapping up and Reporting Results 99
5.4 Case Studies 100
5.4.1 Marie Boucher Social Network 100
5.4.2 Lineages at VAST 102
5.4.3 Feedback from Practitioners 104
5.5 Discussion 105
5.5.1 Limitations 106
5.5.2 Performance 106
5.6 Conclusion 107

6 Conclusion **109**
6.1 Summary 109
6.2 Discussion 110
6.3 Perspectives 113
6.4 Conclusion 116

List of Figures

1.1 Business contract originated from Nantes (France) during the 17th century. See [67] for more detail of the historian process to analyze her sources. 4
1.2 Marriage, partnership, trading, banking, and real estate networks of the powerful families of Florence from [179]. We can see the central position in the network of the Medici Family. 6
1.3 Abstraction of the VA process. It is characterized by continuous interactions between the data, visualizations, models, and knowledge. Image from [127]. . . 7
1.4 Node-link diagram of a medieval social network of peasants, produced with a force-directed layout, commonly used in SNA softwares. The picture is note very informative and only reveal a semblance of community structure. Image from [33]. 8

2.1 Categorization of visual variables which can be used to represent network data, resulting in many different network representations. Image from [25]. 13
2.2 Anscombe quartet. The four datasets have the same descriptive statistics (average, variance, correlation coefficient) but very different structures. Image from [9]. 14
2.3 TULIP software is designed for application-independent network visual analytics [12]. The view shows a dataset among multiple interactive coordinated views. Users can also apply data mining algorithms on the data to extract interesting patterns. 16
2.4 Correspondence letters of Benjamin Franklin and his close relationships, visualized with a map and a histogram, accessible online on the republic of letter website [72]. 20
2.5 Moreno's original sociogram of a class of first grades from [167] (left). The diagram shows 21 boys (triangles) and 14 girls (circles). The same sociogram using modern practices generated from Gephi from [95] (right). The color encodes the number of incoming connections. 22
2.6 All possible graphlets of size 2 to 5 for undirected graphs 23
2.7 Cicero's personal communication network represented with a node-link diagram. Image from [5] 26
2.8 Different criteria are proposed to enhance node-link diagram readability. Image from [134] 28
2.9 NodeTrix system showing a scientific collaboration social network with clusters. Each cluster is represented as a matrix, Image from [111]. 29
2.10 Vistorian interface [209] used to explore a historical social network of business trades in the 17th century, with a coordinated node-link diagram and a matrice view. 30

3.1 HSNA workflow is split into five steps: textual sources acquisition, digitization, annotation, network creation, and network visualization/analysis. Practitioners typically have to do back and forth during the process. I list potential pitfalls for each step. 38
3.2 Three properties essential to VA systems supporting the social historians workflow: *traceability*, *document reality*, and *simplicity*. 41
3.3 bipartite multivariate dynamic network modeling for two cases of marriage acts of example #3. Some marriage acts mention the parents of the spouses, which is a relationship different than the marriage in itself. This case can be modeled using a document model (a) or an event model (c) by splitting the document into several different event nodes. The other case refers to documents that do not mention the parents (b), and in that case, the network represents both the documents and the events with the same model. M: Marriage, H: Husband, W: Wife, T: Witness, (H/W)(M/F): Husband/Wife Mother/Father. Yellow links refer to parenting mentions/relationships. 49

4.1 The ComBiNet system was used to compare two subgroups of a social network of contracts from [55], extracted with dynamic visual queries. (A) and (B) show the two visual queries created by the user in the query panel using an interactive node-link diagram editor (V6), dynamic query widgets (V7), and the equivalent Cypher script (V8). The right part shows ComBiNet's global interface in *comparison* mode: (V1) Network visualization panel, (V2) Map of the geolocalized nodes, (V3) Table of persons, (V4) Graph measures comparison, (V5) Attribute distribution plots, and (V9) Provenance tree. The two visual queries on the left, translated into Cypher queries below, select the "Menafoglio" family on the left, and the "Zo" family on the right, along with their construction contracts and close collaborators. 61
4.2 ComBiNet interface wreal-timeith the dataset of collaboration #1. The user selected the _*year* attribute, showing the distribution of document years with a histogram (bottom right), and coloring the documents node on the bipartite view (left) and map view simultaneously (right). 63
4.3 All link creation possibilities: Any link type (left), one selected link type, here guarantor (middle left), the union of several link types (middle right), several links with different types (right) 64
4.4 Visual queries created to answer questions 2 and 6 of our collaboration #1. (a) The visual query retrieves individuals who are mutually guarantors to each other in separate construction contracts. (b) The two visual queries retrieve the documents—along with the signatories—of Torino (*Turin* in french) (left) and of Torino's surroundings (*Turin Territoire* and *Piemont*) (right). 65
4.5 Widget designs for the different attribute types: checkboxes for categorical attributes (top), text input for nominal attributes (middle), and a double slider for numerical attributes (bottom). The categorical attribute example shows the inputs letting users create new constraints for other attributes and other nodes. 66

4.6 Results of question 2 of collaboration #1: (a) shows a subset of the table view with every occurrence of the pattern found. (b) shows the summary panel, with the graph measures and the attributes view with the *origin* attribute selected and the Sankey option checked. It allows us to see the attribute distribution of the persons included in the pattern and see if there is a relationship between persons who are mutually guarantors and their origin. 67

4.7 Two ways of showing the distribution of "type chantier" (type of works), a categorical attribute with three possible values "*rel*igious", "*mil*itary", and "*civ*ilian". (a) A query matching the contracts made by the same person (*per1*) as an "approbator" (green link to *doc2*) after being a "guarantor" (blue link to *doc1*) using the constraint (`doc2._year > doc1._year`). (b) Stacked bar chart for the matches, the earlier contract (*doc1*), the older contract (*doc2*), and (c) Sankey diagram with the early values on the left and the last on the right. The Sankey diagram reveals the value changes between the two documents: the guarantor who worked initially on religious work switched to military work. . . . 68

4.8 Provenance tree to answer question 2 of collaboration #1: left branch leads to Torino documents (the node is labeled as A) while right branch leads to surrounding documents (the node is labeled as B). The user hovers over one node, revealing a tooltip that shows the visualization of the node's query. . . . 69

4.9 Comparison table of the network measures for Torino subgraph (A) and Torino's surroundings subgraph (B). 70

4.10 Distribution of the type of constructions, the years, and the betweenness centrality for the documents and signatories of Torino (A), Torino's surroundings (B), and the whole graph (top). 71

4.11 Menafoglio (a) and Zo (b) families were retrieved with queries and highlighted in the bipartite node-link and map views. 73

4.12 Attributes distributions plots between the whole graph, the *Menafoglio* family (A), the *Zo* family (B), and $A \cap B$, for the *region,type_chantier, material type.* 74

4.13 Map of the migrations in France which occurred across several generations. . . 75

4.14 Migrations across departments over three generations 75

4.15 Sankey diagrams showing the migration of people in the 18th (left) and 19th (right) centuries, extracted from their birth and death places. 76

4.16 ComBiNet used to request persons appearing as husband, wife, or witness in two marriages that occurred 70 years apart or more. 77

4.17 ComBiNet used for exploring theses of sociology defended in France between 2016 and 2021. The bipartite and map views show an overview of two visions of the network. The user selects the *region* attribute, showing the geographical distribution of the defended thesis. 78

4.18 Sociology thesis dataset explored with ComBiNet. The user constructed a visual query to see if there are symmetrical relationships between thesis directors and reviewers (or jury directors). The *region* attribute is selected with the Sankey option, letting the user see if there are correlations between the regions of the thesis found in this pattern. 79

5.1 Process of traditional clustering (left) and our PK-Clustering approach (middle and right). The output of traditional clustering is a possible clustering, using an algorithm among many choices. The output of PK-Clustering is a clustering supported by algorithms' consensus and validated (fully or partially) according to the user's PK. 86
5.2 Prior Knowledge specification, the user-defined two groups composed of two members. 91
5.3 Red edges represent the prior knowledge matching 93
5.4 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. 95
5.5 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. . . . 97
5.6 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. 97
5.7 The suggestion of extra clusters. The two PK-groups (red and blue) are validated (nodes in the consensus column are all squared). One extra cluster 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. 99
5.8 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. 100
5.9 Two main phases of PK-clustering. On the left, the user has specified the Prior Knowledge (PK) groups (top left) and then reviews the list of algorithms ranked according to how well they match the PK. On the right, the user compared the detailed results of selected algorithms and consolidated the results. From the initial specification of three groups and three people, 4 relevant clusters were obtained with 37 people in total, plus one unclassified node (*Others* group). . . 101
5.10 Summary report of the consolidated knowledge for the Marie Boucher case study. 102
5.11 Computing the Lineages of VAST authors: Prior Knowledge from Alice and results of the clusterings matching it. 103

5.12 Four consolidated groups in the VAST dataset: C North, RVAC, Andrienko, and London 104

6.1 Potential place of VA in the HSNA workflow defined in chapter 3. 109
6.2 Prototype of a document-centered dynamic layout for bipartite multivariate dynamic networks, showcased with construction contracts in Piedmont (see collaboration #1 in chapter 3). The layout can show the mentions of persons encoded with their roles (top) or the documents and their properties (bottom). Here, the *region* attribute is selected, hence coloring construction contracts depending on their locations. 115

List of Tables

2.1 Comparison table of most widely used visualization and analytical tool for SNA. Visualizations: number of different visualization techniques, and layouts. SNA Measures and Models: number of proposed SNA measures and algorithms. Clustering: Number of proposed clustering algorithms. Filtering: Possibilities of filtering according to various criteria. Interaction/Direct Manipulation: Number of possible interaction mechanisms directly applicable to the visualizations. . . 30

3.1 Resulting networks using different models produced by one document of the examples detailed in §3.3.1: co-occurrence, unipartite and bipartite models. The first column shows the partial transcription of real documents (simplification for collaboration #1). Colors represent annotations concerning the persons mentioned, their roles, and their attributes. Underlines refer to information related to the events and which can be encoded as document/event attributes. Only time is represented for simplification, but other attributes would follow the same schema. H: Husband, W: wife, T: Witness, M: Marriage, A_N: Associate, G: Guarantor, Ap: Approbator, C: Construction, F: Father, M: Mother, C: Child. 45

4.1 Tasks to support during exploration, according to our expert collaborators, are split into 3 main high-level tasks. 60

4.2 Comparison of the data model of several VA systems aimed at exploring bipartite social networks. 60

Publications

Publications related to the thesis

- A. Pister, P. Buono, J.-D. Fekete, C. Plaisant, and P. Valdivia, Integrating Prior Knowledge in Mixed-Initiative Social Network Clustering, IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 2, pp. 1775–1785, Feb. 2021, doi:10.1109/TVCG.2020.3030347.

- A. Pister, N. Dufournaud, P. Cristofoli, C. Prieur, J.-D. Fekete, ComBiNet: Visual Query and Comparison of Bipartite Multivariate Dynamic Social Networks, accepted with minor revisions to Computer Graphics Forum, Wiley,

- A. Pister, N. Dufournaud, P. Cristofoli, C. Prieur, J.-D. Fekete, From Historical Documents To Social Network Visualization: Potential Pitfalls and Network Modeling. VIS4DH 2022 - 7th Workshop on Visualization for the Digital Humanities, Oct 2022, Oklahoma City, United States.

- A. Pister, C. Prieur, and J.-D. Fekete. Visual Queries on Bipartite Multivariate Dynamic Social Networks. Poster, EuroVis 2022 - 24th EG Conference on Visualization, Jun 2022, Rome, Italy. doi:10.2312/evp.20221115

Other Publications

- N. Tovanich, A. Pister, G. Richer, P. Valdivia, C. Prieur, J.-D. Fekete, and P. Isenberg. VAST 2020 Contest Challenge: GraphMatchMaker: Visual Analytics for Graph Comparison and Matching. IEEE Computer Graphics and Applications, IEEE, 2021, 42 (4), pp.89 - 102. doi:10.1109/mcg.2021.3091955.

- S. Di Bartolomeo, A. Pister, P. Buono, C. Plaisant, C. Dunne, and J.-D. Fekete. Six Methods for Transforming Layered Hypergraphs to Apply Layered Graph Layout Algorithms. Computer Graphics Forum, Wiley, 2022, EuroVis 2022 - Conference Proceedings, 41 (3), pp.1467-8659. doi:10.1111/cgf.14538.

- G. Richer, A. Pister, M. Abdelaal, J.-D. Fekete, M. Sedlmair, and D. Weiskopf, Scalability in Visualization, submitted to Transactions on Visualization and Computer Graphics, 2022, doi:10.48550/ARXIV.2210.06562

1 Introduction

> "*My claim rests on the assumption that [...] researchers can learn the truth about social processes. At a minimum, they can distinguish between totally inadequate and less inadequate representations of social processes, thus opening the way to increasingly reliable knowledge.*"
>
> **-Charles Tilly, [229]**

The goal of this thesis is to characterize and produce visual analytics tools that can support social historians conducting research on their sources—particularly when using network methods—with a focus on exploration, analysis, traceability, and usability. *Historical Social Network Analysis (HSNA)* is a method—sometimes referred as a paradigm [244]—followed by social historians to study sociological phenomena through the observation of relationships of actors of the past, modeled into a network. The usage of networks as an abstraction to represent and study social relationships—such as friendships, kinship, or business ties—grew in popularity in the last 40 years [85, 224] and constitute a powerful metaphor, especially in our time when many of our digital connections and interactions use an explicit network structure[1]. This approach has first been formalized in sociology under the term *Social Network Analysis (SNA)* [85] and is now widely used in anthropology [163], geography [], and history [129]. Historians leverage historical documents—which are at the core of their profession [140]—to extract relationships between actors of interest that they model with networks constructed from nodes and links that respectively represent actors (often persons) and relationships (like kinship). Using social network visualization techniques and leveraging network measures and computations, they can then test hypotheses they have and gain insight on the structural aspect of the relational phenomena they are studying [129, 242]. This approach has been followed successfully to study various subjects such as kinship [103], entrepreneurship [200], maritime routes [141], political power [179], political oppositions [178], and persecution [160]. Yet, history is considered by many as a literary and qualitative science, and many critics emerged from the history community concerning quantitative and network methods [100, 125, 146, 149], pointing to problems such as the leading to trivial conclusions, anachronisms, simplifications, and mismatches between network and historical concepts. Moreover, quantitative and network analysis are complex processes, and demand many efforts in data collection, encoding, modification, and processing before being able to make efficient observations. This thesis considers the whole workflow of social historians to better support it with visual analytics.

Social historians have to take many annotation (sometimes called encoding) and modeling decisions, concerning *what* to model from their sources into a network, and *how* to model it [53, 66], i.e., should the information of interest be represented as a node, a link, an attribute, or not reflected in the network at all, and what format should be used. Practically, they typically use

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

ad hoc processing and analysis scripts to transform historical documents to analyzable networks, which is time-consuming, sometimes to end up with trivial or hard to interpret results [6]. Still, HSNA led to many highly regarded studies with thorough conclusions, such as the study of families of power in Florence by Padgett and Ansell where they explained the rise of the Medici family through its central position in the economical, political, and trading networks of powerful families [179] or Gribaudi and Blum work on the social and professional shift during the 19th century in France [98].

The usage of visualization to graphically display networks is common in SNA[2] as it allows to unfold the structure of networks to the eyes, thus letting social scientists confirm hypotheses they had when collecting and exploring their data as well as gaining new insight through the discovery of interesting patterns and trends [54]. Images of networks also constitute an efficient mean of communication, especially in scientific productions [84]. Many visualization techniques and softwares have thus been developed since the beginnings of SNA, but most popular tools are usually not designed for historians specifically, meaning that they do not regard on the provenance and process leading to the network, and focus on analysis aspects only. Moreover, they typically enforce simple network models without proposing exploration mechanisms, beyond allowing to look at the network structure and computed measures. As a result, many HSNA studies show a plot of their network and describe it qualitatively, often by identifying the central actors—sometimes with the help of centrality—but do not go beyond that [147]. *In this thesis, I therefore investigate how visualization can support social historians in their work, first during the pre-analysis process and secondly during the analysis step, with the right levels of expressiveness, usability, and traceability.*

1.1 Social History and Historical Social Network Analysis

Social history has continuously evolved since its beginning in the 1930s, especially with the rise of quantitative and network methods based on the development of computer science during the end of the 20th century. If these computer-supported methods are now widely used in history [129, 183], they attracted many criticism from the start—some are which still relevant.

We can trace back the birth of social history with the formation of the "Annales School" in the 1930s, where historians gained interest in socio-economic questions and started to rely heavily on the exhaustive extraction and analysis of historical documents coming from archives [27, 189]. Beforehand, history was mainly political and event-centered, as the majority of work consisted in narrating and characterizing specific events—such as wars and diplomatic alliances—while eliciting their causes and consequences, and describing the lives of historic figures, such as sovereigns [189]. Social history shifted the focus by aiming to link together sociological, economical, and political issues and by placing individuals at the center of these questions [228]. Later on in the 1960s, with the development of computer science, historians

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

started to use quantitative methods to analyze data extracted from historical documents and make conclusions grounded in statistical results, in various subjects such as demographics [110] and economics [94]. Around the same time, the use and study of networks started to become popular in various disciplines to study real-world relational phenomena based on mathematical computations and measures, especially in sociology and anthropology [41]. A network is an abstraction based on graph theory concepts which can be used to model phenomena based on relationships (called links) between entities (called nodes).

Sociologists appropriated this concept to model social relationships between agents of interest, allowing them to study the sociological structure of groups of interest—such as families, institutions, and companies—and concepts like friendship, oppression, and diffusion using real world observation and mathematical computations. This SNA approach allows analysts to ground results in formal network measures and metrics based on real observations instead of relying on traditional social categories such as age, job, and gender [85]. This shift in the object of study from traditional social classes and aggregates to the observation of relationships of individuals remind the microhistory movement [90] which theorized that following the life of single individuals and small groups enable the making of higher level conclusions about the social structures they live in. Social historians followed this tradition and started to appropriate network concepts to study relational aspects of the past and formalized it under the term Historical Network Research or Historical Social Network Analysis [242]. However, historians do not have the possibility to run surveys or directly observe interactions of the past and are thus constrained by the information contained in historical documents they find in archives. These documents can be anything mentioning social relationships between actors of interest, such as marriage acts, birth certificates, census, migration acts, business transactions, and journals. After selecting a corpus of documents, they typically read and inspect in depth several documents while taking notes to have a deeper insight on the content of the sources, which allow them to start eliciting hypotheses. Following this exploration phase, they manually annotate each document and encode the desired information—the mention of persons and their social relationship in the case of a network analysis. This is a long and tedious process that can result in small to large networks that they analyze using network measures to make conclusions on the structure of social groups or social behaviour of individual of interests. Figure 1.1 shows for example an original business document of the 17th century from Nantes (France). The historian have to inspect these documents in depth, extract useful information, and cross-reference the sources to do her quantitative analysis afterwards. The investigation and reading of the historical documents is therefore an exploratory process, where historians start to generate sociological hypotheses from the continuous extraction of insight and revelations of this process, similarly to grounded theory [92]. Once they finalised a network, they can test their hypotheses using qualitative or quantitative methods—based on statistical and network measures. Lemercier and Zalc write "Although history is not an exact science, counting, comparing, classifying, and modeling are nevertheless useful methods for measuring our degree of doubt or certainty, making our hypotheses explicit, and evaluating the influence of a phenomenon." [147] Social historians, therefore, have hypotheses about their subject of study, that they can back up or refute with the help of quantitative and network results, in a way similar to the competing hypotheses workflow

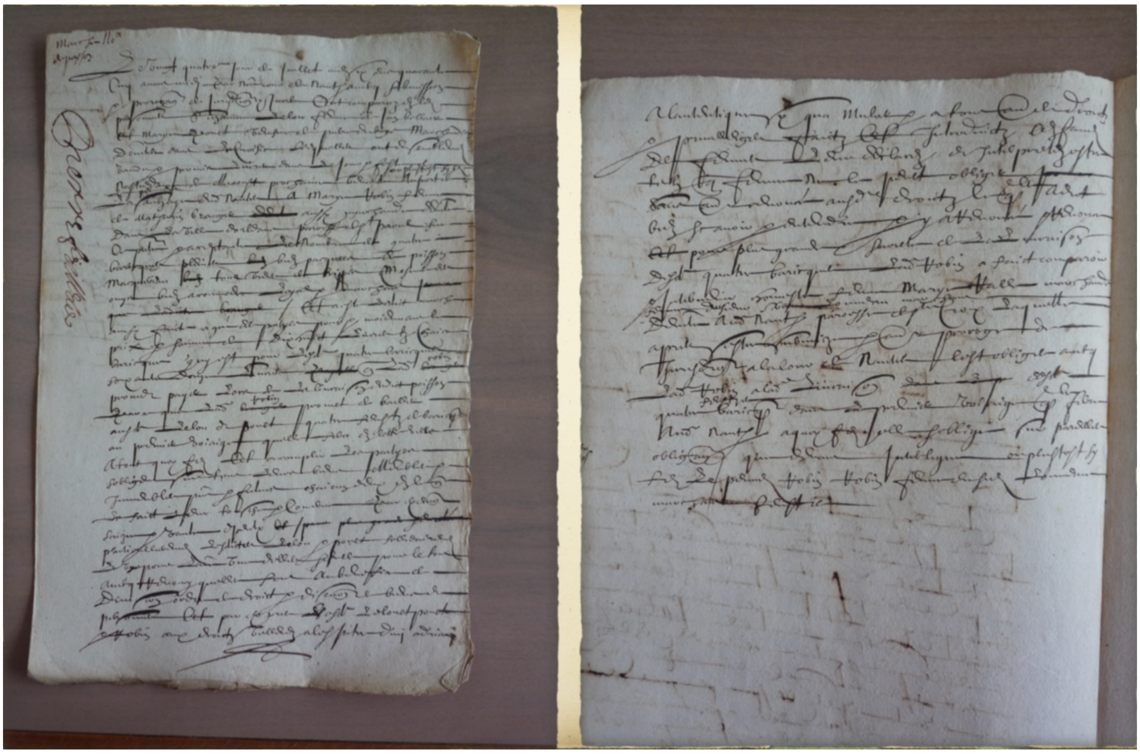


Figure 1.1 – Business contract originated from Nantes (France) during the 17th century. See [67] for more detail of the historian process to analyze her sources.

of Intelligence Analysis [62]. By pointing to evidence supporting or refuting hypotheses, they can give insight into the level of the plausibility of different claims.

1.2 Visualization and Visual Analytics

Visualization has been said to be a central part in the development of SNA [84, 248]—as it the case for many scientific fields[3]. Social scientists now widely use visual and analytical tools to unfold their network structure, allowing them to confirm or deny hypotheses, or follow exploration analysis.

Visualization is the process of displaying data visually to leverage the human visual system and enhance cognition to gain insight into data [44]. Using visual abstractions (such as size, color, and position) to display abstract data allows us to rapidly see structure and patterns otherwise hidden in raw text and numbers. As data keeps growing in size with time due to the increase of hardware and storage capabilities, visualization is a powerful tool to gain insight into the underlying structure of various complex datasets.

[3] the historian Alfred Crosby went as far as claiming that visualization is one of the two factors—with measurement—which led to the development of modern science [57].

Visualization has traditionally been used for confirmatory and communication purposes, particularly in empirical sciences [215]. By showing data visually, analysts are able to confirm or refute hypotheses and communicate their findings in scientific productions.

However, visualization can also be used for exploration, which can help to understand the underlying structure of data and generate new hypotheses. Tukey defined this process as *exploratory data analysis* in the 1960s [233], as a procedure to gain insight into the structure of the data by identifying outliers, trends, and patterns with the usage of visualization and statistical measures. Social network visualization is used for communication of findings in the field, but is also often following this exploration process as showing the network visually allows social scientists to reveal the structure of their data. As freeman writes "Images of social networks have provided investigators with new insights about network structures and have helped them to communicate those insights to others" [84]. Social scientists very often represent their data using node-link diagrams, that we find in every production of reference in the field [36, 143, 224, 241].

Figure 1.2 shows a node-link representation of the network constructed by Padgett and Ansell in their work on the Medici. At that time, diagrams were often drawn by hand, practice which have now been replaced by automatic layout algorithms. Most visual software for SNA such as Gephi [18], Pajek [171], NodeXl [218], or Ucinet [122] are based on this representation, and allow an exploration of the data with the help of basic interaction mechanisms and the computation of network measures. The detection of patterns and trends can also be facilitated with automatic methods coming from data mining and machine learning fields, directly implemented in the visual analysis loop. This coupling of visual exploration and automatic data mining algorithms has been coined as Visual Analytics (VA) and is defined as the process of using interactive visualizations, transformations, and models of the data in an interactive analysis workflow to create knowledge [127].

Figure 1.3 illustrates the schematic process of VA: the coupling of visualization and data mining models operated by the user through interaction lead to the generation of knowledge "extracted" from the data. If most widely used visual interface for HSNA do not yet provide complex interactions or high data mining capabilities, more recent tools are oriented towards VA, as the combination of automatic knowledge extraction with interaction and exploration can be a powerful support for social scientists to gain insight on the structure of their network, especially that the data they study keep growing in size and complexity [124].

1.3 Visual Analytics Supported Historical Network Research

Most visual tools for SNA are designed for the analysis of already curated networks, without taking into account the context in which those networks have been produced, where they come from, and the workflow that led to their creation. Moreover, many practitioners have trouble using current computer-supported tools, due to misconception in their encoding and modeling process or usability problems [6]. VA should therefore support social historians in the entirety of their process, with a focus on usability and simplicity.

Currently, social historians spend a long time in their data acquisition, processing, encoding,

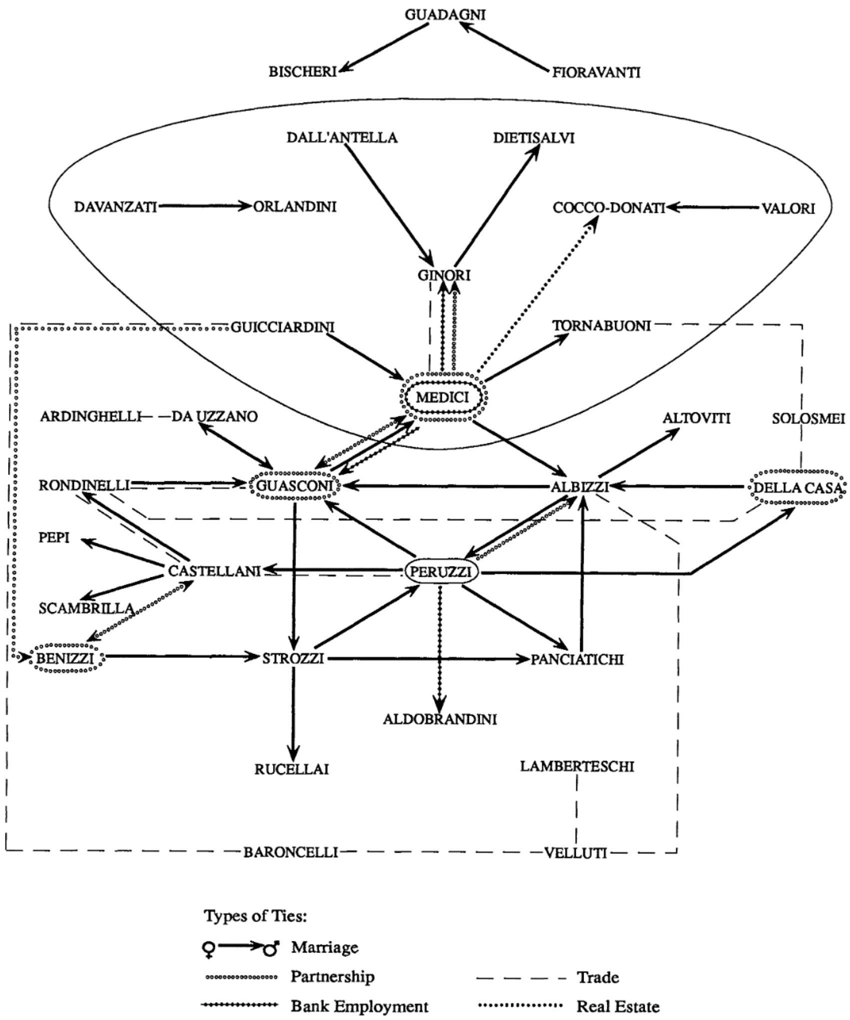


Figure 1.2 – Marriage, partnership, trading, banking, and real estate networks of the powerful families of Florence from [179]. We can see the central position in the network of the Medici Family.

and modeling steps which lead them to the construction of a network [67, 148]. They typically visualize and analyze their network at the end of this process, first to verify hypotheses they formulated during the inspection of their sources, then to gain a better view of the structure of the network, allowing them to potentially generate new hypotheses [146]. However, research showed that all the steps preceding the analysis can introduce errors and misconceptions, especially since social scientists are often not trained in computer science and data science [6, 147]. Social scientists usually visualize their network using SNA tools like Gephi, Pajek, and NodeXl which encompass basic interactions, node-link visualization, SNA measure computations, and clustering algorithms. Once they visualize their data, they typically notice errors and inconsistencies in the data, such as duplication of the same entities, merging of different entities, or geolocation errors [6, 64].

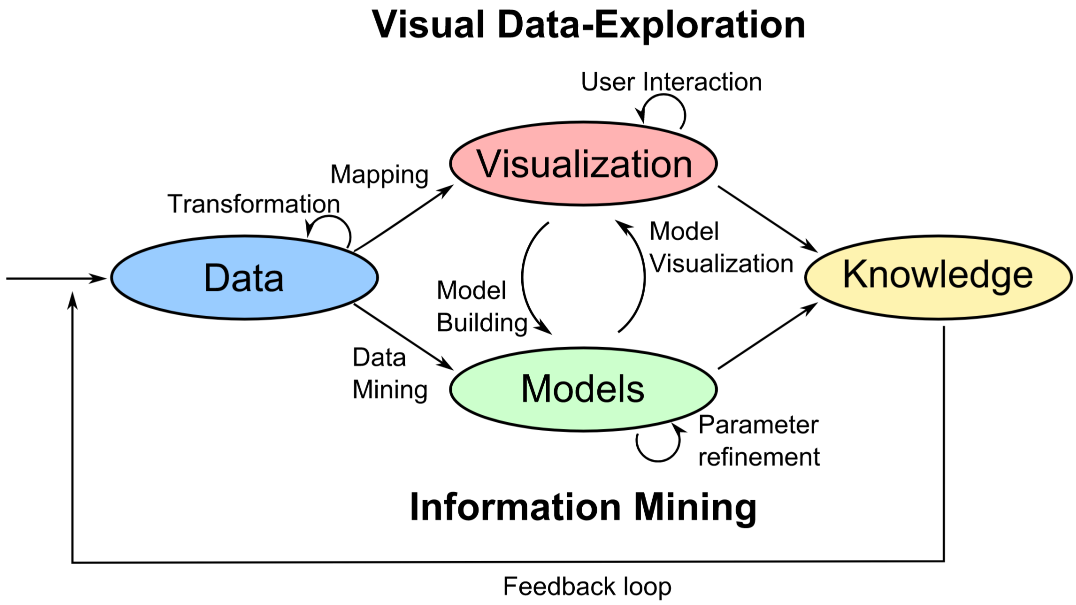


Figure 1.3 – Abstraction of the VA process. It is characterized by continuous interactions between the data, visualizations, models, and knowledge. Image from [127].

Practitioners also have to decide on a network model [53] (see §2.3.4 for more details) when encoding their documents, which sometimes do not match the final analysis goals. Simple models typically oversimplify the relationships contained in the sources [146] and too complicated models are hard to manipulate [177]. They, therefore, have to go back and forth between the visualization software and the encoding process which can be tedious, especially since it can be complicated to trace back the entities of the data model back to the original documents for correction. VA tools that encompass the whole process of social historians should therefore be beneficial for the flow of their work and could help detect and correct errors or analysis plans way before the visualization of a finalised network. Proposing how to design such interfaces with proof-of-concepts is one of the goals of this thesis.

Furthermore, several historians highlighted the fact that many social history studies leveraging network methods simply use networks in a metaphorical sense, in what Rollinger calls "soft SNA" or "informal network research" [196]. Such studies typically show one—or a couple—node-link diagram(s) which they describe with qualitative terms [147] to refer to the global structure of the network (dense, parse, connected, etc.), the place of actors (central, distant), or interesting patterns (cliques, bridges, communities). In case of dense networks, such descriptions become obsolete, as diagrams start to look like what have been called a "spaghetti monster" [49, 147] i.e., an unreadable image due to the high level of cluttering. Figure 1.4 shows for example a medieval social network of peasants proximity relationships between 1250 and 1350, extracted from agrarian contracts. The graphic does not convey much information, especially that the links represent a constructed notion of proximity without indicating the types of relationships the individuals were mentioned in the contracts.

Figure 1.4 – Node-link diagram of a medieval social network of peasants, produced with a force-directed layout, commonly used in SNA softwares. The picture is note very informative and only reveal a semblance of community structure. Image from [33].

The lack of use of network analytical methods—which are numerous in modern SNA softwares[4]—have been in part explained by "math anxiety" [181]: it takes long effort to learn the mathematical concepts behind network measures and algorithms, and their relationships to sociological concepts [196], especially for practitioners without formal computer science and mathematical training. My claim is that current HSNA tools do not support social scientists

[4]See for example the long technical manuals of Pajek [171] and Ucinet [123]

enough in their analysis due to 1) the lack of interaction, direct manipulation, and exploration mechanisms in current interfaces and 2) the lack of network measures and algorithm interpretations and explainability. For example, clustering algorithms are often included in such systems, letting social scientists partition networks into groups, but many algorithms exist in the literature, potentially giving diverse results. Scientists often run several algorithms until finding a satisfying enough partition, which can bias the result of an analysis [186]. Usability and traceability of the results are therefore primordial in VA interfaces aimed at supporting social historians in their analysis.

VA could therefore help social historians in their use of network methods, first by providing guidance and continuous feedback on the inspection, encoding, modification, and modeling process from the sources, and by providing complex exploration and analysis mechanisms supported by data mining capabilities. For this, such interfaces should 1) be simple enough to manipulate, 2) model the original documents and annotations without distortions, and 3) let historians trace back their network entities to the original sources and analytical results in explainable frameworks. In other terms, they should satisfy *simplicity*, *document reality*, and *traceability* principles. I discuss and explain those principles more in depth in chapter 3.

1.4 Contributions and Research Statement

The goal of this thesis is to characterize how VA can support social historians in their HSNA process and present proofs of concepts of tools supporting it. Most social network visualization tools are agnostic to the process of social historians leading to a polished network, even though it has a high impact on the network model and structure. Using visualization only at the end of the process often reveals potential errors, inconsistencies, or mismatches between the network model and analysis goals [6]. Moreover, due to lack of usability and interaction mechanisms, social historians often simply visualize statically their network and partially describe their structure, leading to conclusion which would have been easier to reach with simpler methods [75]. VA could therefore 1) assist social historians in their overall workflow, starting at the documents' acquisition to the final analysis step, with the help of data mining and interaction mechanisms in the data acquisition, encoding, modeling steps, and 2) provide exploration and analysis mechanisms to answer complex historical questions, beyond simply plotting the network with a node-link diagram.

The goal of this thesis is hence to give answers to the high-level question "How can VA support social historians in their entire HSNA process?". To answer this question, I first characterize the HSNA process from start to finish from discussions and collaborations with social historians, with the goal of identifying pitfalls that regularly arise and characterizing social historians' needs. From this, I give answers and directions—illustrated by proof-of-concepts—to three questions concerning the modeling aspect of HSNA and how VA and automatic tools can support social historians in different parts of their process, while satisfying *traceability*, *document reality*, and *simplicity* properties:

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

- **Q2**: What representations and interactions would allow social historians answer complex historical questions—with a focus on usability?
- **Q3**: How to design VA tools and interactions that leverage algorithmic power but keep historians in control of their analyses and biases?

In chapter 3, I start by describing the HSNA workflow and identify recurring pitfalls we encountered in our collaborations with historians and answer **Q1** by proposing a network model for modeling historical documents. In the following chapter 4, I give answers to **Q2** by providing a VA interface to explore bipartite multivariate dynamic networks, with queries and comparison interactions with the aim of letting historians find errors easily, transform their network data, answer their questions, and generate interesting hypotheses. Finally, in chapter 5, I propose PK-Clustering, a mixed-initiative clustering technique for social scientists based on their prior knowledge, algorithmic consensus, and traceability of results, as a concrete example of a system addressing **Q3**.

2 Related Work

Contents

2.1 Visualization 12

2.1.1 Information Visualization 12

2.1.2 Visual Analytics 15

2.2 Quantitative Social History 16

2.2.1 History, Social History, and Methodology 16

2.2.2 Quantitative History 18

2.2.3 Digital Humanities 19

2.3 Historical Social Network Analysis 20

2.3.1 Sociometry to SNA 21

2.3.2 Methods and Measures 23

2.3.3 Historical Social Network Analysis 25

2.3.4 Network Modeling 25

2.4 Social Network Visualization 27

2.4.1 Graph Drawing 28

2.4.2 Social Network Visual Analytics 29

Social historians rely on textual historical documents to study social groups through their structures and socio-economic characteristics in societies of the past [40, 228]. They read and analyze documents they can find from a period and subject of interest, and make their conclusions through deep inspection and cross-referencing of the information they found. Several methods have been developed in history to extract and analyze the information contained in the documents in a methodical way [229], based on qualitative or quantitative methods—among which HSNA is now widely popular [196]. HSNA is a method consisting in modeling the relational information mentioned in the documents—such as family, business, or friendship ties—in a network, to be able to characterize and explain social behaviors through the description of the network's structure [129, 242]. This approach is directly inspired by SNA, which is a well-known method that sociologists theorized to understand and describe real-world social relationships modeled as networks [85, 207]. Historians appropriated this method, by extracting relationships from historical documents. The specifics of HSNA in contrast to its sociology counterpart is, therefore, the modeling of the network from the historical documents—which are at the core of the historical work [189]—and the integration of the temporal dimension which is often disregarded in traditional SNA but central in history. Once they successfully constructed a network—which is a long and tedious process—they typically use network measures and visualization techniques to confirm or generate new hypotheses [146]. Visualization let them unfold the structure of their data, revealing potentially interesting social patterns between actors and groups of actors. Analytics and visualization systems for SNA typically allow the exploration of such data with the help of interaction, network measures, and data mining capabilities such as clustering directly implemented in the interfaces. Yet, most HSNA studies only give a qualitative description of their network—which Rollinger calls "soft" or "informal" network research [196]—probably due to usability and formalism issues [6]. The coupling of visualization and data mining through interaction to support the generation of knowledge has been described as VA and can therefore provide support to social historians for their network construction, but also to go beyond simple qualitative description of their data. In this chapter, I first present a general overview of the field of visualization in §2.1 to share its utility and potential for social history. Then, I present the social history discipline with its use of quantitative methods in §2.2, before describing in depth how network analysis has been applied in the field in §2.3. Finally, I present in §2.4 how visualization and VA have been used in the context of HSNA, along with the most popular systems currently used by social scientists and their limitations.

2.1 Visualization

Visualization is often defined as "the use of computer-supported, interactive, visual representations of data to amplify cognition" [44]. Graphically displaying data allows us to leverage our visual system to gain a better acquisition of knowledge, leading to better decision-making, communication, and potential discoveries. The field of visualization can be split into three subdomains: *Scientific visualization* focuses on visualizing continuous physically based data such as weather, astrophysics, and anatomical data, sometimes produced with simulations whereas *Information Visualization* is centered around the visualization of discrete abstract data points, often multidimensional. *Visual Analytics* emerged later from Information Visualization by mixing data mining and more complex analysis process with traditional information visualization displays. I focus in this thesis on the two latter branches of visualization, as social scientists can use both information visualization and VA systems to gain insight into the structure of the networks they are studying.

2.1.1 Information Visualization

Information Visualization focuses on displaying abstract data to amplify cognition and gain insight into real-world phenomena [44]. History is filled with classical examples of visual data displays which helped understand better specific events, such as Minard's map of Napoleon's march in Russia [87], or Snow's dot map of cholera cases in London which showed the proximity between street pumps and cholera infections [219]. If several examples of information visualization can be found thorough history, it mainly developed as a scientific field in the 1960s with Tukey's work on data analysis and visualization [232] and Bertin's publication of *Semiology of graphics* [25].

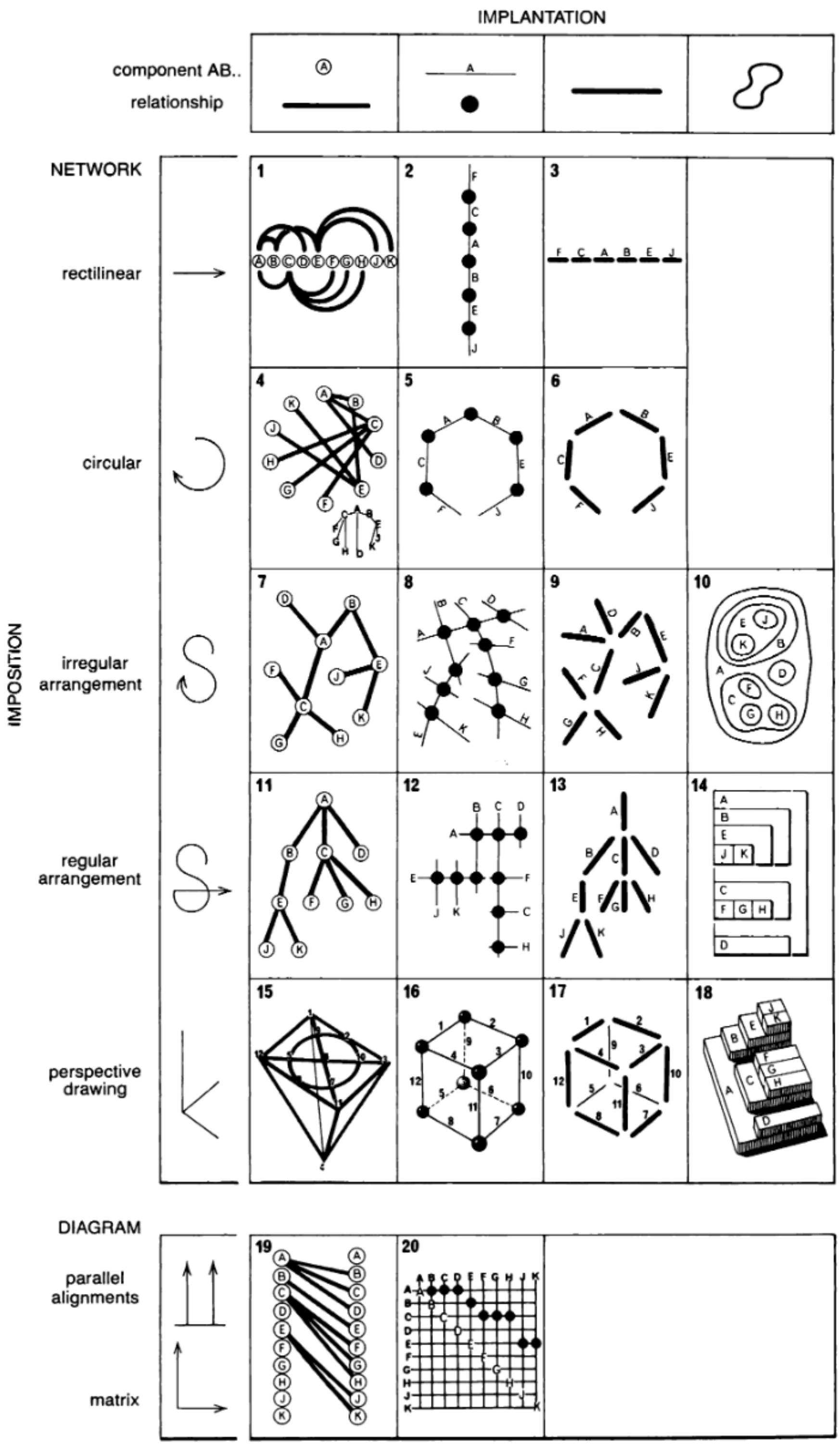


Figure 2.1 – Categorization of visual variables which can be used to represent network data, resulting in many different network representations. Image from [25].

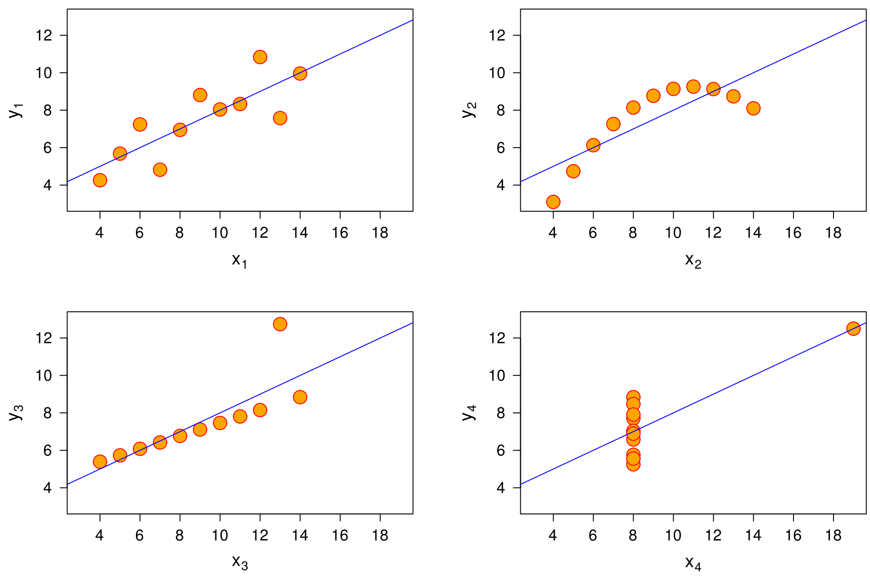


Figure 2.2 – Anscombe quartet. The four datasets have the same descriptive statistics (average, variance, correlation coefficient) but very different structures. Image from [9].

In this foundational work, Bertin described and organized the different visual elements usable in graphical information displays, and linked them to data features and relations types. An illustration of this work of categorization for network data is illustrated in Figure 2.1. Michael Friendly writes "To some, this appeared to do for graphics what Mendeleev had done for the organization of the chemical elements" [88]. The development of computer science and the rise of hardware capabilities at the same time created a big need for data visualization. The amount of data stored increased exponentially [112] and descriptive statistics were not enough to understand the underlying structure of the amount and diversity of produced data. Visualization, leveraging the human visual system, enabled to rapidly see the hidden structure of a dataset and detect interesting and unexpected patterns very often unseen with classical statistical methods. One classical illustration of this is Anscombe's quartet [9] which consists of four datasets of 11 points in $\mathbb{R}^2$ with the same statistical measures (mean, variance, correlation coefficient, etc.) but with very different structures, that are immediately revealed when plotting the data. The four datasets are illustrated in Figure 2.2.

A large number of visualization techniques emerged to make sense of the diversity of data produced, such as multidimensional, temporal, spatial, or network data [213]. Instead of using taxonomies classifying graphics into categories such as histograms, pie charts, and stream graphs, some theorized how to describe graphics in a more systematic and structural way. In 1993, Wilkinson extended Bertin's work and developed the *Grammar of Graphics* [246] as a way to describe the deep structure unifying every possible graphic, thus allowing to characterize and create graphics using common terms and rules. In this framework, a graphic can be defined as a function of six components: *data* (a set of data points and attributes from a dataset), *transformations* (statistical operations which modify the original data, e.g., mean and rank transformations), *scales* (e.g., linear and log scales), *coordinate systems* (e.g., cartesian and polar coordinate systems), *elements* (graphical marks such as rectangular or circular marks, and their aesthetics, e.g., color, and size), and *guides* (additional information such as axes and legend). Many well-known visualization toolkits are now based on this framework, such as vega [204] and ggplot [245], as it enables a greater expressiveness and reusability for graphic creation. Visualization allows to gain insight into the structure of a given dataset and has traditionally been used for confirmation and communication purposes [215], for example, to verify hypotheses on empirical sciences, and later on to communicate findings, first to scientific peers, and nowadays to broader audiences for example through the means of data journalism [35].

2.1.2 Visual Analytics

VA consists of the coupling of visualization and data mining techniques to better support users in their knowledge generation process through continuous interaction with the data and statistical models [227]. It draws inspiration from exploratory visualization, interaction, and data mining. The process of exploratory visualization to gain new insights on the general structure of the data and potentially generate new hypotheses has been characterized by Tukey in 1960 as *exploratory data analysis* [233]. It consists in trying to characterize the structure of a dataset with the help of continuous visualization and statistical measurements of different dimensions of the data. Visual exploration is enhanced by direct manipulation interfaces through interaction and usually follows the information-seeking mantra formalized by Shneiderman: "Overview first, zoom and filter, then details-on-demand" [213]. It allows users to first have a visual overview of the data and get an idea of its overall structure, to then change the point of focus to highlight interesting patterns with the help of filtering, querying, sorting, and zooming mechanisms. As the average size of datasets keeps growing, exploratory tools are often needed to make sense of large datasets and generate pertinent hypotheses.

More recent visual exploration interfaces also incorporate automatic analytical tools along with graphical displays, letting users apply data mining algorithms directly in the exploratory loop. This coupling of visualization and analytical methods such as data mining has been defined as VA and is still a very active research field. Keim et al. define it as "a combination of automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data" [127].

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

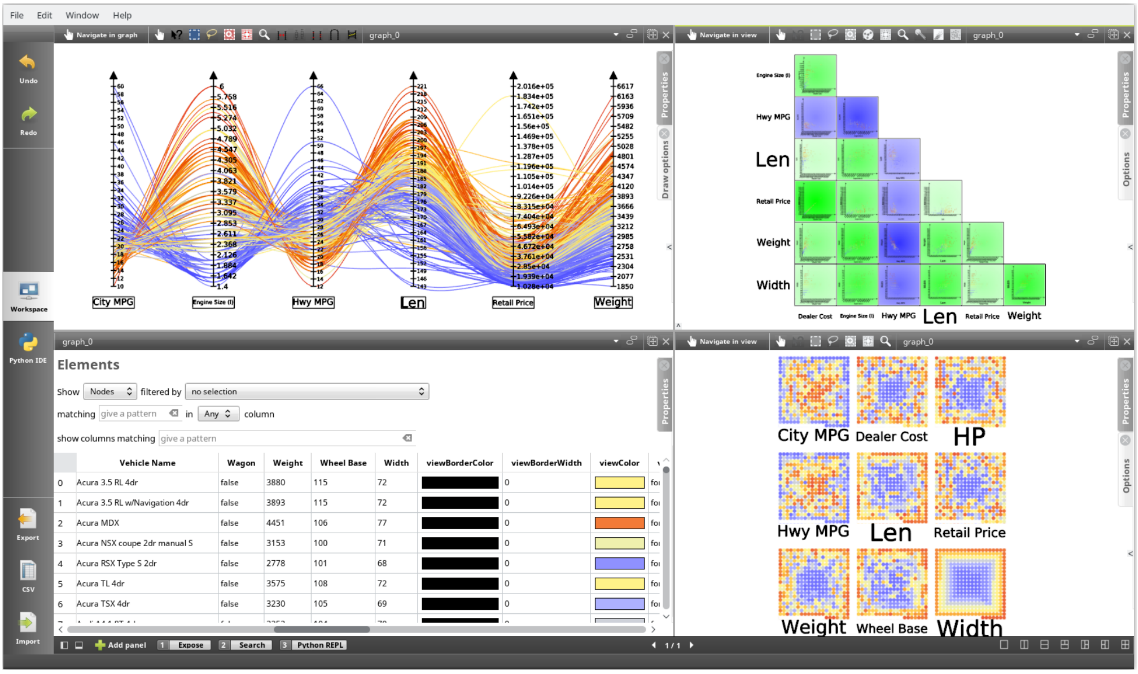


Figure 2.3 – TULIP software is designed for application-independent network visual analytics [12]. The view shows a dataset among multiple interactive coordinated views. Users can also apply data mining algorithms on the data to extract interesting patterns.

2.2 Quantitative Social History

Social History is a branch of history that aims at studying socio-economic aspects of past societies, with a focus on groups instead of specific individuals only. Charles Tilly writes that its goal is to "(i) documenting large structural changes, (2) reconstructing the experiences of ordinary people in the course of those changes, and (3) connecting the two" [228]. If the purpose of social history remained the same across time, methods and formalisms have evolved since its beginning in the 1930s. Specifically, the rise of computer science led to the development of quantitative history methods in the 1960s—now often referred to as *digital humanities*—which brought new ways of grounding results in formalisms and quantitative models, instead of solely relying on qualitative inspection of historical documents [106]. I discuss in this section the evolution of social history from the context of its beginning to the use of more recent quantitative approaches.

2.2.1 History, Social History, and Methodology

The concept of history is hard to define as its practice and codes highly evolved through time. Prost writes "history is what historians do. The discipline called history is not an eternal essence, a Platonic idea. It is a reality that is itself historical, i.e. situated in time and space, carried out by men who call themselves historians and are recognized as such, received as history by various publics [189]." Retrospectively, history of a given time can thus be characterized by the different historical work produced at that time. Nevertheless, History can be characterized as the collection and study of historical documents to study and describe the past. As Langlois and Seignobos write, "The search for and the collection of documents is thus a part, logically the first and most important part, of the historian's craft" [140]. History emerged as a field with its own rules, conventions, and journals in the 1880s from faculties of letters, to counterbalance previous history works which were judged as too "literary" [175]. At that time and until now, two facets characterize the field, which are sometimes overlapping: one is political whereas the other one is methodological. The former aspect of history serves to create a shared story for countries and a sense of unity among citizens. Antoine Prost says for example that "it's through history than France thinks itself" [189]. The latter aspect of history constitutes a methodology to describe the past through methodical inspection of historical sources, with the aim of inferring dated facts about the past and trying to minimize possible bias. Historical documents are thus at the core of the work of historians and having to cite historical documents and previous peers' work for new claims is primordial to be considered rigorous History work. However, methodological and epistemological facets (how historians should read and analyze their sources, how to cite them, what to report/not report, and what is the status of proof) of History have not been studied and discussed for a long time, until the end of the 1980s [189]. Some historians were interested in historiography [43], but none were going to philosophical and epistemological debates on the History discipline. For Lucien Fèbvre, philosophizing was even constituting a "capital crime" [77].

Retrospectively, we can still observe shifts in the objects of study of historians through time, and their relation to sources. History was at first mainly event-centered and was focusing on characterizing central figures of the past like rulers and artists or shedding light on central events like wars or political crises. This narrative approach to history has been criticized for its open interpretation of historical documents, which can introduce bias from the authors [34]. In the 1930s, March Bloch and Lucien Fèbvre detached from traditional history by creating the "Annales School" (École des Annales) which aimed at placing humans as a component of a broader sociological, political, and economic system with influences on each other [40]. They strongly advised exhaustively searching from archives, to ground historical results in documents, texts, and numbers. This new way of studying past events and societies became successful in a profession in crisis, by bringing a new lens of study on various societal subjects more grounded in sources and with better intelligibility. This school of thought can be seen as one of the biggest milestones for social history, which focuses on the socio-economical aspects of societies and their changes through time, rather than an event-centric view of History. For example, in his thesis, Ernest Labrousse—a well-known figure in social history—tries to describe and explain the economic crisis of France at the end of the "Ancien Régime"[1] through the evolution of the economic power of different social groups such as farmers, workers, property owners etcinstead of solely describing memorable facts about the period [138]. Social history continued to evolve since the 1930s, introducing new methods and concepts, but always with the goal to describe

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

periods and historical facts through a sociological lens and with a strong focus on sources and traceability.

2.2.2 Quantitative History

With the development of statistical methods and computer science, quantitative approaches to history emerged in the 1960s with the goal of analyzing numeric data directly extracted from historical documents. Economists led this first wave of quantification by studying past events using economical concepts and data. This approach called "new economic history" or "cliometrics" was popularized by Fogel's study on the economic impact of the development of railroads in America [80] and Fogel and Engerman's controversial work on the economy of slavery [82]. In the latter study, they extracted numbers of a sample of 5000 bills of slave sales from New Orleans to support the controversial claims that slavery was economically viable and that slaves had a decent material life, which brought up heated debate among the scientific community and the broad audience [243]. Despite the controversy, These kinds of approaches rapidly started to be used in other related domains such as demography, social history, and political history, sometimes rebranded as "new social history" and "new political history" [148]. Using computer-supported methods, historians were able to store data extracted from historical documents and make conclusions based on computational methods such as regression and statistical testing. Many saw the future of social sciences in computer programming, as Le Roy-Ladurie who wrote in 1968 "The historian of tomorrow will be a programmer, or he will not exist" [147].

However, quantitative methods started to be criticized in the 1980s with a wave of disillusionment, for several reasons. Stone was the first to raise his voice in 1979, after participating himself in several of those ambitious projects: "It is just those projects that have been the most lavishly funded, the most ambitious in the assembly of vast quantities of data by armies of paid researchers, the most scientifically processed by the very latest in computer technology, the most mathematically sophisticated in presentation, which have so far turned out to be the most disappointing" [222]. First, many researchers of this first wave dispensed themselves with source criticism, leading to simplification, anachronisms—such as using modern analytical categories and indices like the GDP—and taking the numeric data from historical documents as objective. These problems could be in part explained by the fact that the work process was highly divided, meaning that the people analyzing the data did not necessarily inspect and read the original historical documents in depth. Indeed, "new history" projects often relied on a high division of labor among researchers, assistants, and students who operated with punch card operators [139], since extracting the data from raw documents and uploading it to computers—which were shared among whole departments—was very time-consuming at that time. Secondly, the popularity of these methods made practitioners forget about the many biases inherent to statistics, such as the sampling bias, or the fact that historical data is essentially incomplete data. This resulted in the computation of long data series and aggregates which were sometimes nonsensical given the gaps in the sources [147]. Finally, many historians raised their voices against the study of long-term trends instead of focusing on specific events and individuals. They challenged aggregation procedures and their assumptions, trying to go back to a more complex history by pointing out that phenomena have to be studied and understood through several scales [231]. Indeed, computing correlations and aggregates at a national level greatly simplify complex phenomena and misses specific group and individual related behaviors. Still, if their adoption remains slow and sometimes criticized among historians, quantitative methods provide tools to store, explore, and analyze historical documents systematically if used appropriately (i.e. not trying to bias the analysis, and not losing the trace of the original sources), especially that those methods highly evolved since the 1960s.

2.2.3 Digital Humanities

Digital Humanities is sometimes described as the second wave of computational social sciences [147]. The term has gained popularity since the 2010s and refers to "research and teaching taking place at the intersection of digital technologies and humanities. Digital Humanities aims to produce and use applications and models that make possible new kinds of teaching and research, both in the humanities and in computer science (and its allied technologies). Digital Humanities also studies the impact of these techniques on cultural heritage, memory institutions, libraries, archives and digital culture" [226]. If the first wave of computational social sciences focused a lot on statistical methods such as regression models, correlation testing, and descriptive measures (mean, median, and variance) to make conclusions, digital humanities focuses also on the use of digital tools for exploration, teaching, and communication of humanities concepts and data, leveraging design, infographics, and interactive systems [39]. In the context of historical research, the term *digital history* has been coined as "an approach to examining and representing the past that works with the new communication technologies of the computer, the Internet network, and software systems. On one level, digital history is an open arena of scholarly production and communication, encompassing the development of new course materials and scholarly data collections. On another, it is a methodological approach framed by the hypertextual power of these technologies to make, define, query, and annotate associations in the human record of the past. To do digital history, then, is to create a framework, an ontology, through the technology for people to experience, read, and follow an argument about a historical problem" [1]. Research that label itself as digital history pivots around the curation and digitization of historical archives, the identification of historical concepts through computational and exploration methods, and also their communication to the general audience through digital technologies. Many Digital History projects are thus multidisciplinary by essence and involve several teams of researchers, such as the *Mapping the Republic of Letters* project which consisted of digitizing, storing, and exploring letters of scholars across the world in the 17th and 18th centuries, in a common hub and using shared visualization tools [72]. It resulted in the elaboration of curated datasets and visualizations concerning the correspondence of various scholars such as Voltaire, Benjamin Franklin (see Figure 2.4), and John Locke, accessible in the same place by researchers and the general audience. With modern technologies and infrastructures, it also becomes possible to study large historical databases—often labeled under the term "big data"—as with the *Venice Time Machine* project [124] which aims at digitizing and analyzing thousands of documents from the archives of Venice to understand the political, geographical, and sociological dynamics of the cities across generations and centuries. Yet, some historians raised concern about this type of project, fearing that it could rapidly bring the same type of issues encountered during the first wave of quantification, especially for big projects

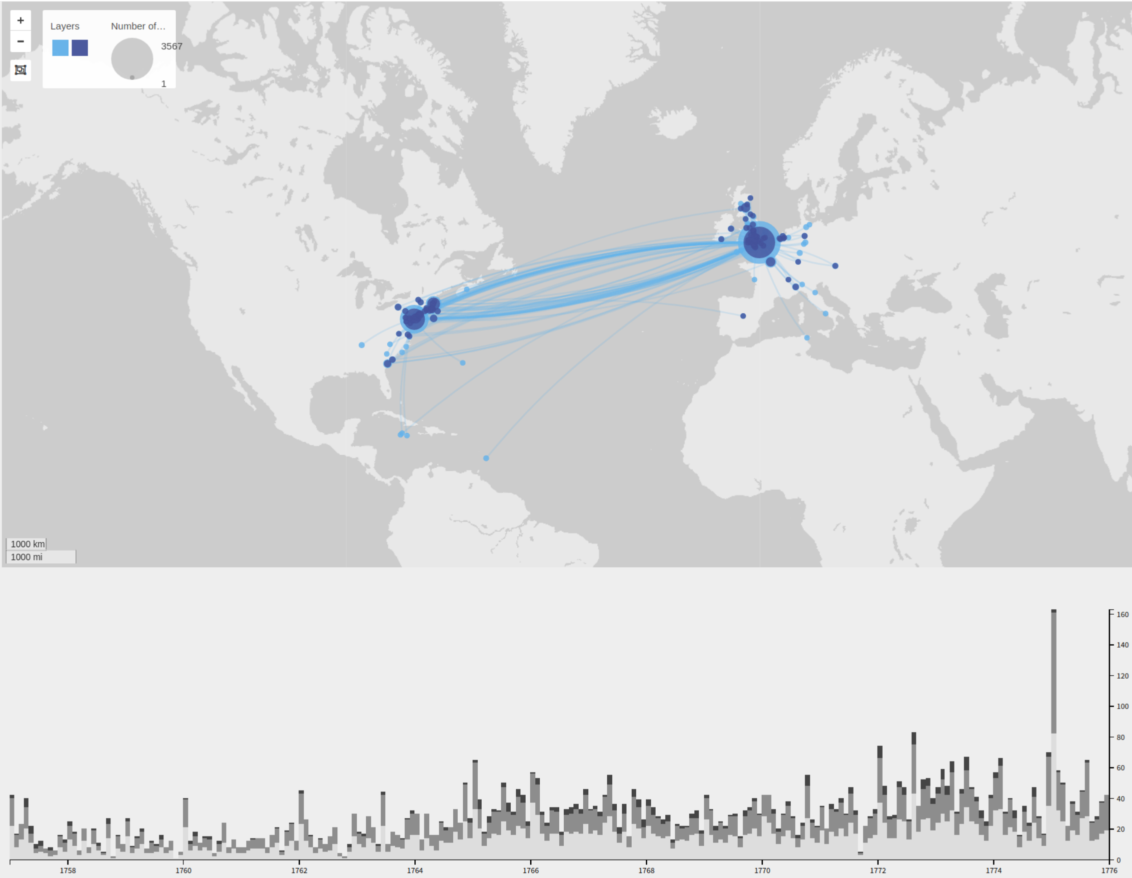


Figure 2.4 – Correspondence letters of Benjamin Franklin and his close relationships, visualized with a map and a histogram, accessible online on the republic of letter website [72].

involving many actors and highly ambitious goals [147]. Guildi and Aritage went as far as criticizing the decrease of interest of historians working in archives [100], since many historical dataset are now directly available online through web technologies and archives portal.

Many projects which claim themselves as digital history also leverage new methods compared to the 1960s and 1970s, such as the use of network methods and concepts [4]. Examples are the *Viral Texts* [52] and *Living with Machines* [11] projects which respectively study nineteenth-century newspapers and the industrial revolution by translating their sources into analyzable networks. I discuss more in detail the related work of network analysis for historical research in §2.3.

2.3 Historical Social Network Analysis

Historians started to use network analysis to study relational structures and phenomena of past societies in the 1980s, using similar methods developed by sociologists under the label of

SNA. SNA can be defined as an "*approach grounded in the intuitive notion that the patterning of social ties in which actors are embedded has important consequences for those actors. Network analysts, then, seek to uncover various kinds of patterns. And they try to determine the conditions under which those patterns arise and to discover their consequences*" [85]. the use of networks emerged in response to traditional sociology methods using pre-defined taxonomies and social categories to understand and explain sociological behaviors and phenomena, which could introduce bias [207]. By modeling real observed social relationships and interactions with networks and by using mathematical and statistical methods to study those, sociologists have been able to explain sociological phenomena and describe sociological interactions through their direct observation modeled as networks. SNA is now a well-praised methodology in sociology and has been extended to historical research to study relational concepts such as kinship, friendships, and business, through the characterization of groups of past societies. Social historians leverage their documents to extract relationships between entities—often persons—that they model into networks. Leveraging network measures and visualization, they can make conclusions through structural observations of such networks.

2.3.1 Sociometry to SNA

One of Sociology's main goals is to study social relationships between individuals and find recurrent patterns and structures allowing us to generalize on how social relations operate, and what are the social specificity of specific groups and individuals [207]. Traditional methods try to answer those questions using classical social classifications such as age, social status, profession, and gender, typically collected from surveys and interviews. Criticism pointed out that this type of division is often partially biased and comes from predefined categories which are not always grounded in reality [85] and that using random sampling of individuals with such methods remove them from their sociological context. The sociologist Allen Barton wrote in this regard "For the last thirty years, empirical social research has been dominated by the sample survey. But as usually practiced, using random sampling of individuals, the survey is a sociological meatgrinder, tearing the individual from his social context and guaranteeing that nobody in the study interacts with anyone else in it" [17]. Sociometry is considered one of the bases of SNA and had the goal of redefining social categories through the lens of real social interactions and ties between persons, which sociologists wanted to observe in real conditions. It is in the 1930s that Moreno started to develop this new method by trying to depict real social interactions as a way to understand how groups and organizations were socially structured [168]. He developed sociograms to visually show friendships between people with the help of circles representing persons and lines modeling friendships. Figure 2.5 shows one of Moreno's original sociograms to depict friendships in a class of first grades (left). Sociometry tremendously helped disseminate the metaphor of networks to model and understand social structures and phenomena. It was during the 1960s that sociologists and anthropologists took these concepts further and formalized SNA using graphs[2] and mathematical methods [41, 85],

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

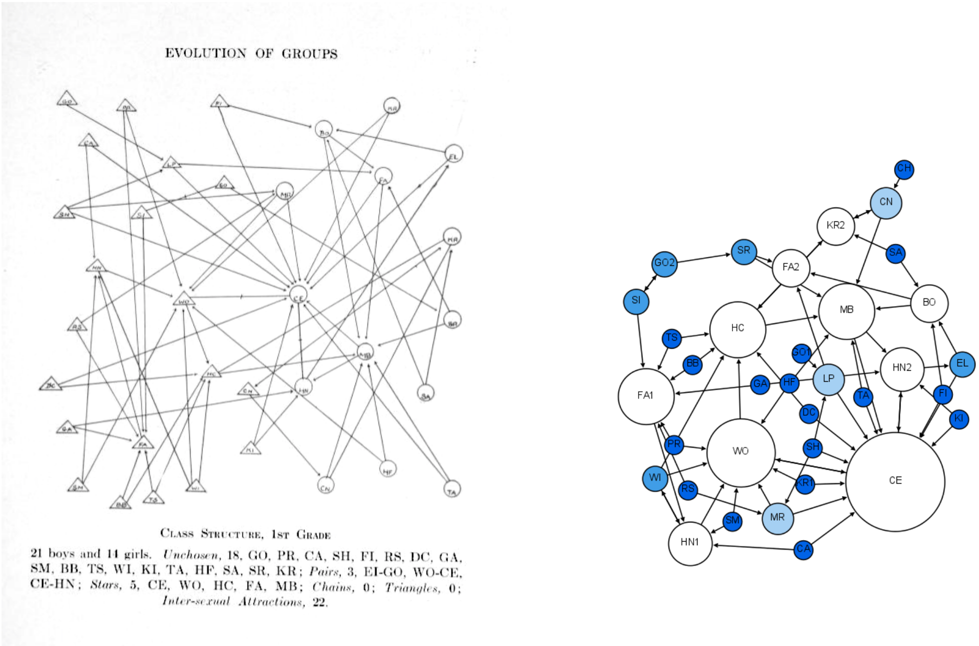


Figure 2.5 – Moreno's original sociogram of a class of first grades from [167] (left). The diagram shows 21 boys (triangles) and 14 girls (circles). The same sociogram using modern practices generated from Gephi from [95] (right). The color encodes the number of incoming connections.

following the emergence of Graph Theory studies in the 1950s by Mathematicians such as Erdős [73]. Sociologists already had structural theories of social phenomena, and they rapidly saw the potential of graphs to model social relationships between actors of interest. Typically, a graph is noted

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

with V a set of vertices representing the actors of interest—typically persons—and $E \subseteq V^2$ a set of edges modeling social relationships. This simple model which does not take into account the diversity and extent of social relationships still allows the characterization of the sociological structure of groups and institutions—which is the primary focus of SNA [85,207]. More complex network models have been proposed with time to better take into account concrete properties of social relationships. I discuss those more in depth in §2.3.4.

Graph theory brought a panoply of concepts and methods to characterize the structure of networks, that sociologists such as Coleman started to codify to use in a sociology setting [48]. The use of network measures let sociologists explain social phenomena through the formal description of real observations of relationships modeled as networks. I describe commonly used methods and measures in the following subsection.

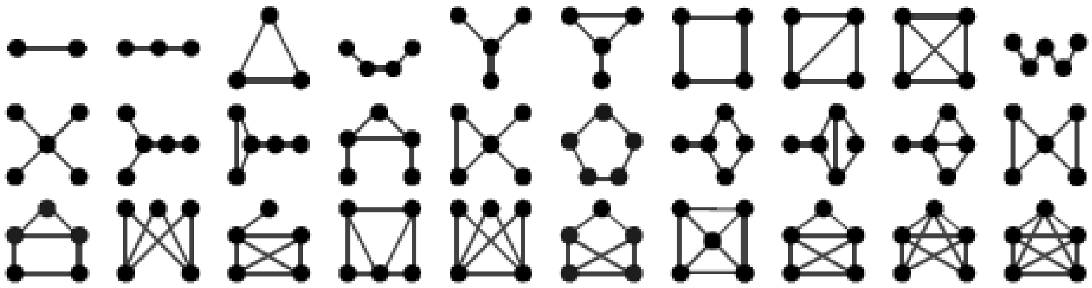


Figure 2.6 – All possible graphlets of size 2 to 5 for undirected graphs

2.3.2 Methods and Measures

Many measures and algorithms have been proposed in the network science and SNA literatures to characterize the structure of simple networks as defined in Equation 2.1 and relate it to social behaviors and phenomena [207, 224]. Network measures are either global or local, which allows one to either make high-level conclusions on the general structure of social relationships constituting the network or individual behaviors. Widely used global measures include for example the density and the diameter, which give insight into the sparsity of the network and how distant on average are two random pairs of nodes. Conversely, local measures give information on the structural position of a node compared to the rest of the network. Centrality—probably the most used local measure—allows to formally compute a measure of how important or central are individuals in the network [172]. As defining what an important node is ambiguous, several types of centrality have been proposed such as the degree, betweenness, and closeness centrality, which respectively measure the number of connections, how nodes connect to different groups, and how close are the nodes compared to the rest of the network.

More generally, sociologists aim at identifying recurring patterns of sociability between actors and linking them to other behaviors, measures, or qualitative knowledge. These patterns can for example be small unconnected components, cliques, or bow-tie structures [241]. Groups of nodes similarly located (central or distant) and having similar shapes are sometimes referred to as "structurally equivalent" [146]. Instead of observing complex shapes, network scientists have also been interested in studying relationships at the lowest possible scale, i.e., observing relations between sets of 2 and 3 nodes at once, also called dyads and triads [241]. This reflects Simmel's formal sociology, where he already referred to dyads and triads as the primal form of sociability [217]. More recently, graphlet analysis extended this concept to every pattern of N-entities [190].

Graphlets are defined as small connected *induced*, *non-isomorphic* subgraphs composing any network [164]. In an *induced* subgraph, two vertices linked in the original graph remain linked in the subgraph. For instance, if the original graph is a triangle  we can only induce the simple edge or triangle  subgraph (graphlet). The path of length 2 has all vertices of the original graph but misses an edge and is, therefore, not a possible graphlet.

Figure 2.6 shows all graphlets of size 2 to 5, for undirected networks. Graphlets counting shows that graphlets are not found in a uniform distribution in social networks [46], thus revealing that social networks do not have the same structure that random networks. Precisely, entities in real-world networks tend to agglomerate into groups (also called *clusters*) where entities in the same groups interact more between them than with entities from other groups [91]. From a sociological perspective, it means that people tend to interact and socialize in groups and interact more rarely with other people from outside groups. These groups are often referred to as *communities*, and many algorithms have been proposed to find these automatically [83].

However, network concepts, measures, and algorithms have not been used only to study groups, organizations, and societies, but also to focus on separate specific individuals. Indeed, two distinct methodologies emerged through the history of SNA: the structuralists and the school of Manchester [75, 85, 156].

Structuralists are interested in observing the relational structures and patterns forming a network, to make parallels between them and the social behaviors of actors in real life [143]. Accordingly, sociologists in this school usually study organizations and specific groups—such as institutions, companies, and families—and want to explain their functioning through the description of the internal shapes and structures of the networks. Thus, they try to construct networks that exhaustively model all the interactions between the actors constituting the groups, as missing links would misrepresent the reality of interactions.

In contrast, the school of Manchester constituted by anthropologists focuses on studying specific individuals and all their interactions in the different facets of their lives and through time. They typically want to explain certain behaviors and social characteristics of individuals by their relationships and interactions in all their complexity and highlight the influence of different social aspects between them in one's life. One famous example is Mayer's study on austral African rural migrants going to cities [157] where he showed that the integration of urban mores and customs was directly correlated to the persons' relationships networks in the city. Xhosa[3] people still interacting with rural people of their village in the city were less changing their customs. This school of thought typically relies on the concept of ego and multiplex networks [75]. Ego networks are networks modeling all the direct relations of one central node—in this case, a person—including the relations existing between the persons of this small network. They typically try to model the different types of relationships of a person, like their family, work, and friendship ties and study them through time. By studying the ego network structure of someone, sociologists of this school try to leverage explanations on other social aspects of the persons like their social status, job, and gender. It is also common to compare several ego networks to make correlations between the social relationships of individuals and other interesting social categories [46].

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

[3] Xhosa people are an ethnic group living in South Africa who talk the Xhosa language.

2.3.3 Historical Social Network Analysis

History started to use concepts and methods from SNA in the 1980s [242] in response to quantitative history, and to develop historical approaches—like *Microstoria* [90]—that focus on the study of individuals and small groups through the lens of their interactions and relationships directly extracted from historical documents. Beforehand, historians were already describing and studying relational structures such as families and organizations with qualitative methods and with classical taxonomies, without necessarily studying the relational aspect of these structures. Network research allowed them to model those relational entities more thoroughly using networks, hence allowing them to make new observations that it was not possible to make without taking into account the relational aspect of these entities [53]. Since then, HSNA—a term coined by C. Wetherell in 1998 [242]—has been applied by historians to study multiple types of relationships, like kinship [102], political mobilization [151], and administrative/economic patronage [170]. If these approaches fall under some of the same criticism as quantitative history [146] like leading to trivial conclusions, it still led to classical work and interesting discoveries, such as the study of the rise of the Medici family in Florence in the 15th century by Padgett & Ansell [179], or Alexander & Danowski study on Cicero's personal communications [5]. In the latter work, they modeled the communication of Cicero into a network using 280 letters written by him between 68 B.C. and 42 B.C. It allowed them to study the relationships between knights and senators—which is a subject of interest in Roman history—and concluded that knight-knight interactions were very rare compared to senator-senator and senator-knight interactions. Cicero communication network is illustrated in Figure 2.7.

Several historians are using and continuously reflecting on HSNA methods [53, 146] which can be very effective to study relational historical phenomena [129]. However, contrary to sociologists and anthropologists who base their networks on direct observations of the real world, historians first have to go through a deep inspection, encoding, and modeling of their sources.

2.3.4 Network Modeling

Constructing a network from historical documents, which can vary tremendously in their formats and structures, is not a trivial task [6]. The most straightforward and well-known approach consists in using simple graphs such as in Equation 2.1, where the nodes refer to the persons mentioned in the documents and links refer to one type of social relationship or a notion of *proximity* constructed from appearance in common documents [33, 146].

This enables to have simple networks to visualize and analyze, but it does not always reflect the sociological complexity of information contained in the documents. HSNA network models have evolved over time to better take into account concrete properties of social networks, such as the importance of actors or relations with weighted networks, multiple relationships with multiplex networks, and dynamics of relationships with dynamic networks.

Weighted networks model the importance of relations, with a weight w attributed to each edge $e = (u, v, w)$, with $u, v \in V$, $e \in E$, and $w > 0$. Multiplex networks allow the modeling of multiple kinds of relationships between actors, such as spouses and witnesses relations for a historical network constructed from marriage acts. In that case, each edge $e = (u, v, d)$ of the

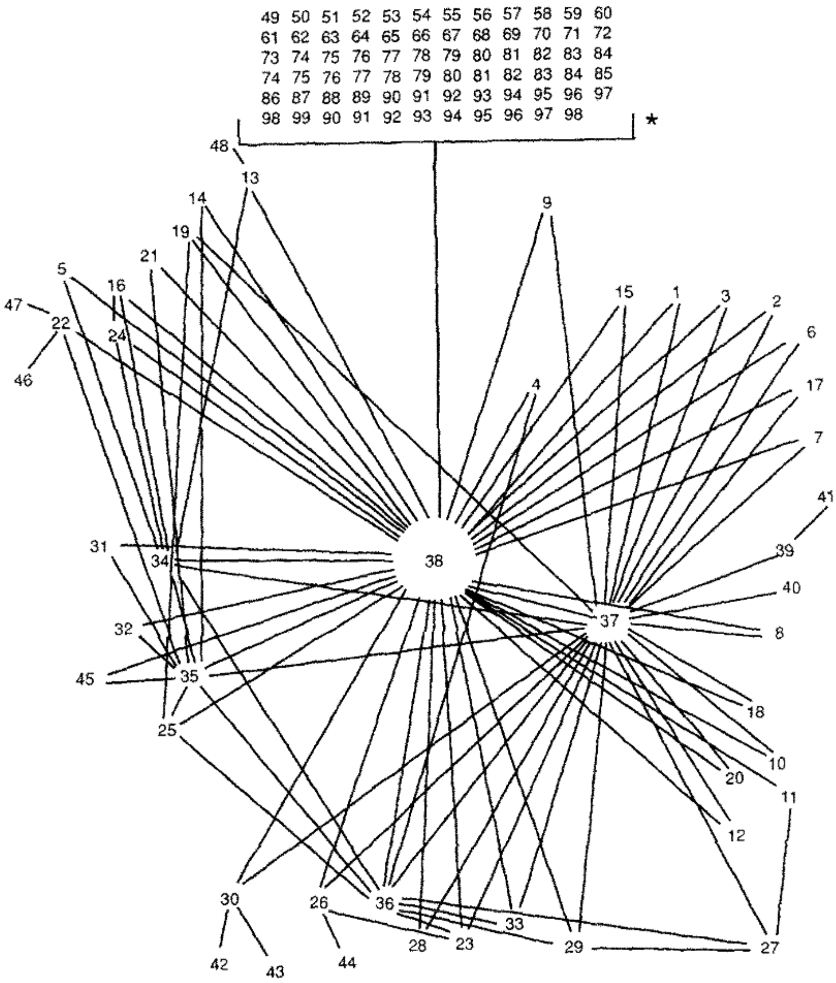


Figure 2.7 – Cicero's personal communication network represented with a node-link diagram. Image from [5]

graph

$$G = (V, E, D) \tag{2.2}$$

have a type $d \in D$ which characterize the relation. In the example of marriages, $D = \{spouse, witness\}$. Most relations extracted from historical documents also often contain time information, which can be modeled into dynamic networks. Many dynamic network models have been proposed [197], depending if the time is encoded in the nodes, the links, or both, and if it is modeled as discrete or interval values. As it is often hard to infer the end of social relationships from the trace of historical documents, I only consider in this thesis models which give a timestamp to either nodes or edges, such that a dynamic graph

$$G = (V, E, T) \tag{2.3}$$

have vertices consisting of tuples (u, t) and/or edges of the form (u, v, t), with $t \in T$.

Bipartite networks have been proposed to model relations between two types of entities, such as organizations and employees where the relations link employees to organizations but not employees to employees or organizations to organizations [30]. Formally, each node of the graph

$$G = (V, E, B) \tag{2.4}$$

have a type $b \in B$, with $card(B) = 2$. For each edge $e = (u, v) \in E$, the types b_u and b_v of u and v are not equal: $b_u \neq b_v$. Many social situations or documents can be modeled in these terms (affiliation lists or co-authoring). Multivariate networks, i.e., graphs, where vertices and edges can be assigned multiple "properties" or "attributes", are less used in SNA. These attributes are often considered secondary, the emphasis of SNA being on the topology, its features, measures, and evolution.

Historians, demographers, sociologists, and anthropologists have also been designing specific data models for their social networks, based on genealogy or more generally kinship [103]. For genealogy, the standard GEDCOM [104] 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 [102] 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.

When creating a network, sociologists and anthropologists can use direct observations of the real world, which is not the case for historians who only have access to biased and partial sources. Indeed, the documents historians inspect are often produced by the political and economical elite of the time, and include the subjective view of the authors, especially for literary sources (letters, journals, books, etc.). Historians, therefore, need to take a critical view of the sources by acknowledging the position of the authors of the documents compared to the rest of the society and include it in the analysis [147]. Furthermore, the partiality of the sources often does not allow to have access to all possible relationship types of individuals. For example, if many formal relations can be extracted from official documents such as marriage acts and censuses, informal relations such as friendships can exist without leaving any written trace [146]. Even for official relationships such as parents and witnesses, there are high chances for missing documents, which do not allow to make too general and finite claims, such as "X is always the case" or "XX is never the case" [5]. Social historians, therefore, have to take into account the partiality and ambiguity of their sources in their analysis, in order to avoid including the bias inherent to their data in their high-level historical conclusions.

2.4 Social Network Visualization

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

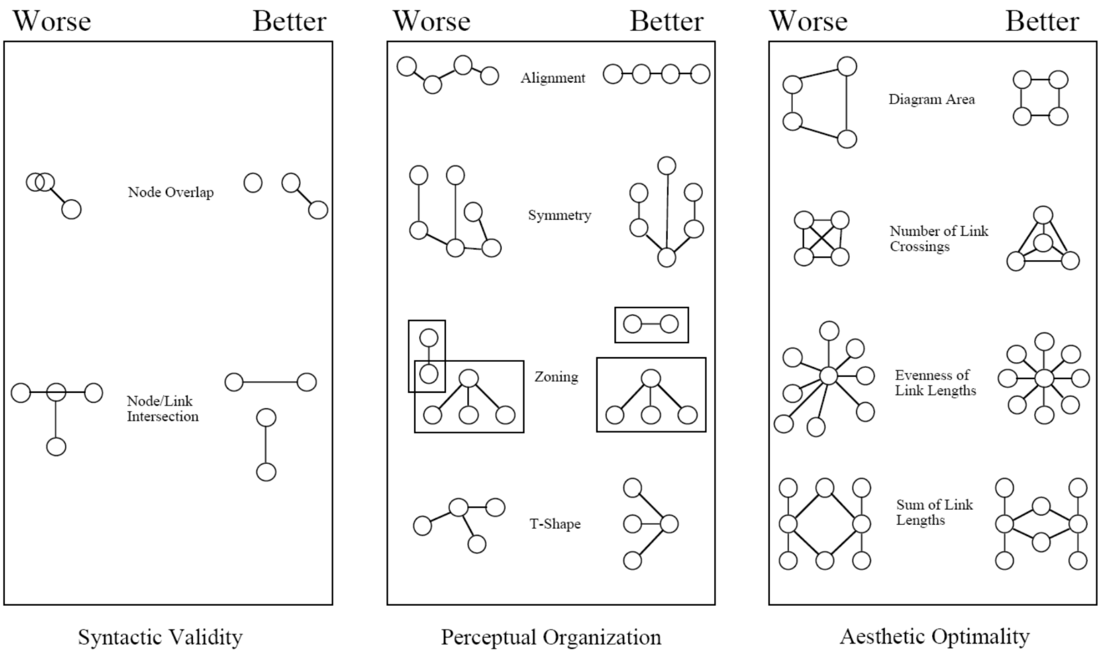


Figure 2.8 – Different criteria are proposed to enhance node-link diagram readability. Image from [134]

a focus on interaction, allowing them to unfold the structure of their data and generte new interesting hypotheses.

2.4.1 Graph Drawing

Sociologists rapidly saw the potential of graphically showing relationships between individuals, to better comprehend the underlying social structure and communicate their findings [84]. Moreno elaborated sociograms to visually show friendships among schoolchildren with circles and lines to respectively show children and friendships ties [167]. This type of representation—commonly called node-link diagram—is the most widely used in social sciences, as it is rapidly understandable and effective for small to medium-sized networks which are predominant in the field. Finding an optimal placement for the nodes is however not that simple as several metrics can be optimized depending on the desired drawing, such as the number of edge crossings, the variance of edge length, orthogonality of edges, etc [54, 134]. Figure 2.8 shows some of these metrics, synthesized by Kosara et al. [134]. In Figure 2.5 we can see the difference in readability between the original manual layout (left) and an automatic one (right). Automatic layouts which aim at optimizing readability metrics give clearer diagrams. The number of edge crossings is often considered the most important measure, but finding a drawing with the optimal number of crossings is an NP-Hard problem, meaning that heuristics are needed for most real-world use cases. A large number of algorithms have been designed such as force-directed ones [20], modeling the nodes as particles that repulse each other and are attracted together when connected with a link using a string analogy. Other visual techniques have been pro-

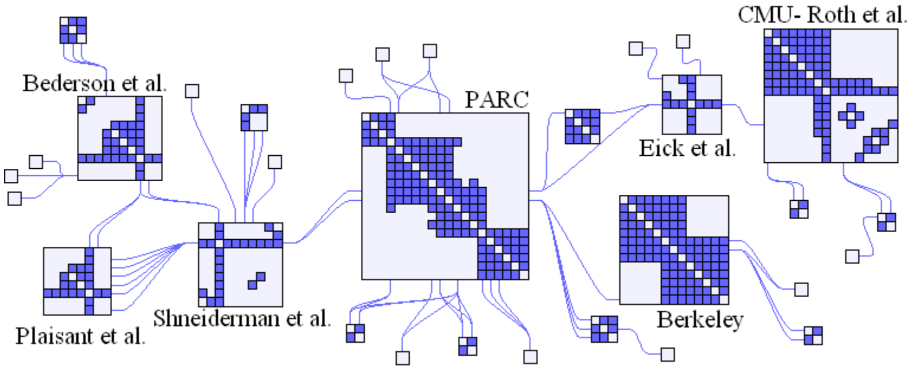


Figure 2.9 – NodeTrix system showing a scientific collaboration social network with clusters. Each cluster is represented as a matrix, Image from [111].

posed to represent networks such as matrices, circular layouts, and arcs, but are less used in social sciences [159]. Still, Matrices have been shown to be more effective than node-link diagrams for several tasks such as finding cluster-related patterns, especially for medium to large networks [2, 89].

As social scientists are using more complex network models such as bipartite or temporal networks, more sophisticated representations are needed. The visualization community developed new representations to visualize other network types such as dynamic hypergraphs with PAOHVis [234], clustered graphs with NodeTrix [111] (illustrated in Figure 2.9), geolocated social networks with the Vistorian [209], and multivariate networks with Juniper [174]. However, these new network representations take time to be adopted by social scientists who rarely use them.

2.4.2 Social Network Visual Analytics

Social scientists use visualization and analytical tools to gain insight on the structure of their finalized network data. The most widely used tools are Gephi [18], Pajek [171], Ucinet [122], and NodeXl [218], which provide node-link diagrams, implementations of network measures, algorithms, and clustering capabilities. Other SNA visualization tools have been proposed in the past such as Visone [22]. However, those tools often have usability issues as they do not include interaction and direct manipulation mechanisms, making the analysis more tedious for social historians. In contrast, the Vistorian [209] let social historians visualize their network with multiple coordinated views (node-link, matrice, arc-diagram, and map), filters and direct manipulation, but do not integrate analytical options. Figure 2.10 shows the Vistorian interface used to explore a historical social network. I propose a qualitative classification of all those tools in 2.1 which rank them according their visualization, SNA measures and modes, clustering, filtering, and interaction capabilities. It illustrates that the most used tools are analytical-oriented (Pajek, Ucinet, Gephi, NodeXl) without proposing many visualization and direct manipulation options, while the Vistorian is an interactive visualization tool without any analytical method integration. None of those systems therefore fully correspond to the VA label [127].

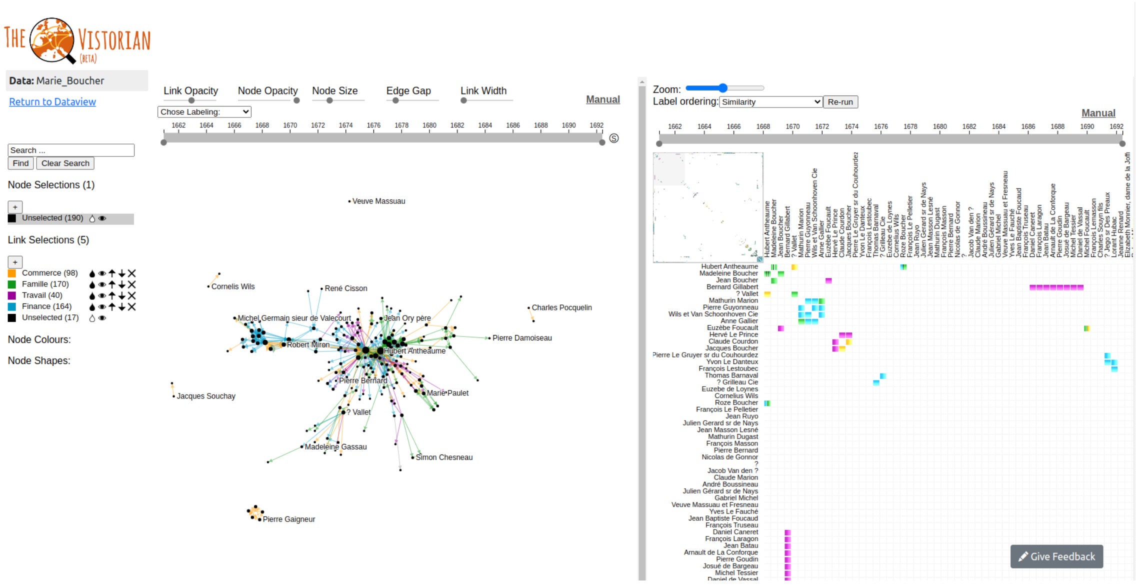


Figure 2.10 – Vistorian interface [209] used to explore a historical social network of business trades in the 17th century, with a coordinated node-link diagram and a matrice view.

	Visualizations	SNA Measures and Models	Clustering	Filtering	Interaction and Direct Manipulation
Pajek [171]	■□□	■■■	■■□	■□□	□□□
Ucinet [122]	■□□	■■□	■□□	□□□	□□□
Gephi [18]	■■□	■□□	■□□	■□□	■□□
NodeXl [218]	■□□	■□□	■■□	■□□	□□□
Vistorian [13]	■■■	□□□	□□□	■■■	■■■

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

If analytical methods such as the computation of network measures, triad computation, and clustering provide a good framework to describe the structure of a network and link it to sociological explanations [207, 241], many social scientists, such as historians, are not trained in computer science and mathematical methods, and, therefore, have trouble to 1) use those methods without guidance and without high usability [6, 196], and 2) interpret their results. This is particularly the case for black-box algorithms such as for clustering tasks: they typically end up trying several algorithms until they stumble upon a satisfactory enough solution [186].

Moreover, preparing and importing the data into visual and analytical software is complicated, as the annotation and network modeling processes have not been globally formalized and every historian use different methods, formats and models. Many users do not succeed in importing their data into those systems without concrete help and guidance [6, 209] due to mismatches with data models, formats, or data inconsistencies (null values, white spaces, etc.). If they succeed in visualizing their data, it often shows them these inconsistencies or errors such as duplications of entities or wrong attribute values. In other cases, they realize the network does not allow them to answer their sociological questions [146]. It leads to continuous back and forth between their analysis process inside the analysis tool they are using, and their annotation/modeling process, to correct errors or modify annotations. Interestingly, the network model choice plays a crucial role, as a simple network model representing only the persons (as is often the case) makes it harder to trace back to the original documents containing the annotations from the network entities. Yet the majority of SNA systems enforce simple network models, making this retroactive process harder.

Some interfaces not primarily designed for social scientists incorporate data models encapsulating document representations, such as Jigsaw [220] which is a VA system using textual documents as a data model, originally developed for intelligence analysis. It allows an analysis of the documents and their mentions of entities (persons, locations, institutions, etc.) through multiple coordinated views. Using such a model allows us to rapidly see errors and inconsistencies in the document annotations that the user can directly correct, while still following complex analyzes.

Finally, more work is still to be done on social network VA tools, to provide more guidance and power to social scientists while doing their analysis, and to help them to do easier back and forth between the annotation, network modeling, correcting, and analysis steps, as errors and inconsistencies can cause high variations in the network structure and hence the analysis results [64].

3 Historical Social Network Analysis Process, Pitfalls, and Network Modeling

Contents

3.1 Context **34**

3.2 Related Work **35**

3.2.1 History Methodology 35

3.2.2 Historian Workflows 36

3.3 Historical Social Network Analysis Workflow **37**

3.3.1 Examples 37

3.3.2 Workflow 38

3.3.3 Visual Analytics Supported Historical Social Network Analysis 40

3.4 Network Modeling and Analysis **42**

3.4.1 Network Models 43

3.4.2 Bipartite Multivariate Dynamic Social Network 46

3.5 Applications **47**

3.6 Discussion **48**

3.7 Conclusion **48**

I describe in this chapter a formalization of the HSNA workflow followed by social historians, to shed light on their process and summarize recurring pitfalls to identify how VA could support them in this workflow. Most HSNA practitioners report on their findings concerning the network they constructed from their sources, but few highlight the process which led to these conclusions from the raw historical documents, even though they have to make several encoding, modification, and modeling decisions that deeply influence the final analysis [6]. Specifically, social historians can model documents and their content through various network models which have been proposed in the literature. I discuss this step in depth as it impacts the annotation and analysis possibilities, and I give an answer to our first research question **Q1** by proposing to model this type of data with bipartite multivariate dynamic networks. This model satisfies *simplicity*, *document reality*, and *traceability* properties, which we define as critical for social history work from our joint collaborations with social historians and current critics of HSNA [72, 146, 148].

This chapter is an updated version of an article presented at the VIS4DH workshop of the IEEE VIS: Visualization & Visual Analytics Conference 2022 and published in IEEE Explore [187]. It was a collaboration with Nicole Dufournaud, Pascal Cristofoli, and my supervisors Christophe Prieur and Jean-Daniel Fekete. I have been leading the discussions, elaboration of concepts, and writing of the paper.

3.1 Context

Tools for social network visualization tend to ignore the context in which the networks are produced, where they come from, and the workflow that led from their origin (e.g., documents, polls, interviews, web scraping) to their network form. Yet, practitioners of social history need to inspect and encode their sources in depth using ad hoc methods to generate a network, and sometimes end with errors or simple networks which do not fit their analysis goals [147]. In this chapter, after describing and characterizing the workflow of HSNA [242] from our collaborations with social historians, I explain why and how effective tools for supporting this process should model social networks in multiple steps to support three essential principles: *traceability*, *document reality*, and *simplicity*. These principles emerged from joint experiences as historians and computer scientists while collaborating on multiple projects, and aim at simplifying the HSNA process while enhancing exploration and analysis options and replicability.

Social historians' goal is to characterize socio-economic phenomena and their dynamics in a restricted period and place of interest and to see how individual people of that time lived through those changes [228]. For this, they rely on historical documents that they inspect in depth to next extract qualitative and quantitative information allowing them to answer their research questions.

To study relational social structures where individuals influence each other such as families, companies, and institutions, historians rely on HSNA by modeling the social relationships between a set of entities—usually individuals—into a network. However, the process leading to the final network from the raw documents is often linear, and it is common that, when visualizing their network, historians spot errors and inconsistencies in the network structure that they could have fixed if the process was more iterative [6]. Moreover, historical documents are often complex, meaning that the annotation and modeling process can be done in many different ways, concerning what to annotate from the documents [148] and how to model the annotation in a network [53]. Several network models have been proposed ranging from simple and specific ones like co-occurrence networks to more general and complex ones such as multilayer networks and knowledge graphs. Simple models allow answering specific questions and are easy to manipulate but are often too simplistic and may distort the information contained in the documents. Moreover, they often break the traceability from the analysis to the original documents, making the communication of findings less reproducible and the process of modifying/correcting annotations complicated. Indeed, errors and mismatches often occur in the annotation process, for example, due to entity disambiguation problems [64]. On the contrary, too complex models are complicated to visualize and analyze, and historians do not always have the tools to create them properly. In this chapter, I answer **Q1** (how to model historical documents into analyzable networks with the right balance between expressivity and simplicity) by proposing to model historical documents as bipartite multivariate dynamic networks, where both persons and documents are modeled as nodes with attributes and the links represent both individuals' mentions in the documents and their social roles in the event witnessed by the documents (such as witness in a marriage act). While this model is simple enough for creation and inspection, it allows tracing back the entities of the network to the original sources for a continuous annotation process and still accurately models the social relationships mentioned in the documents. Historians can therefore use this model to simultaneously find errors and inconsistencies in their annotation process—allowing them easier back and forth between the annotation and analysis steps—while starting a first analysis and exploration of the data to answer their sociological questions. The traceability to the original sources also makes the communication of findings more replicable and transparent.

3.2 Related Work

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

3.2.1 History Methodology

The essence of the historical discipline is based on a critical approach to sources and involves considering peers' work. Traditional approaches to history often focus on the construction of a narrative, without necessarily adopting a systematic and problematized approach to the exploitation of an exhaustive set of historical documents [229]. With the development of social and quantitative history, historians now have a panoply of methods to exhaustively extract quantitative data from their sources and analyze it to ground their results in verifiable claims. Many historians criticized this computational aspect of history [16, 81, 149], pointing out that it would lead to errors and missing the core content of historical sources. However, using quantitative approaches and formalisms is not exclusive to having a deep understanding of the documents and their context, nor building a narrative on top of their quantitative analysis. Good historical work can in fact be described as a combination of the two [125], as Tilly says "Formalisms play their parts in the space between the initial collection of archival material and the final production of narratives. In my own historical research, formalisms figure prominently from early in the ordering of evidence to late in its analysis; [...] As it happens, many other historians rush from sources to reasoned narratives without pausing to employ formalisms, or even to reflect very self-consciously on the logical structure of their arguments, hence on what the evidence should show if their arguments are correct" [229]. Historians have a panoply of methods and formalisms they can leverage to ground their narratives in concrete comparable results, such as serial analysis, tabular analysis, classical statistical treatments, and network analysis.

However, formalisms have to be used wisely and with a critical vision of the documents and their context, so as to not fall into simplifications, anachronisms, and errors which are pertinent critics of quantitative history [147, 148]. Most historical work leverage several methods in the same study to support their claims through different qualitative and quantitative results [183]. The level of the plausibility of a claim increase or decrease depending on if the different evidence point to similar results or not. Similarly, historians often work on small populations or specific individuals—as it is the case with microhistory studies [90]—which can result in complications for generalization. Only after studying several similar individuals or groups, historians are able to generalize and point to exceptions. For example, by comparing several Jewish commercial communities in Europe during the first half of the 18th century, Trivellato has been able to generalize what is common to those groups (they have been trading between them and with outer ethnic groups) and what is specific to each (such as their business strategies) [231].

3.2.2 Historian Workflows

Many quantitative methods and formalisms are available for historians to inspect their sources with the aim of making historical claims. Several textbooks describe and explain to social scientists and students who do not have formal computer science training in what consist these methods (statistical regression, Chi-squared test, network analysis, etc.) and how to practically use those with software and programming language [10, 79]. However, the process leading from the sources to the numeric artifacts (a table, a network, a timeline) has not been described thoroughly in the literature, especially with concrete examples, and is often not presented in scientific publications of concrete use cases. Yet, the process leading from the documents to analyzable data requires social historians to make several annotations, encoding, and modeling decisions, concerning *what* to extract from the source and *how* to encode it. This process is tedious and requires data acquisition, annotation, encoding, and modification with continuous back and forth between the different steps [6]. This is a critical process as it can lead to simplifications, anachronism, distortion, or data that do not allow to answer original or new hypotheses [125, 147]. Lemercier et al. give guidelines on how to encode information from historical documents to prevent introducing bias, by having a critical view of the documents [148]. They emphasize the importance of the input phase of research and advise copying the first documents by hand while characterizing them in the most exhaustive and factual way, without imposing categorization. This explorative step lets historians familiarize themselves with the content of the document, leading to a better view of what to encode to answer their research questions and sometimes to the formulation of new hypotheses. For example, in their project on the social and geographical trajectories of Jews in Lubartów [252], a village in Poland, the team encoded the mean of writing inside the register documents (pen, pencil, ink, etc.) they were inspecting. This information allowed them to conclude that the inscription "expelled" written in pencil was probably added during World War II by Germans to denote exported Jews in the extermination camps. When applying network analysis, historians often create specific person-to-person networks which allow them to answer precise research questions, but often lose this type of document-related information. Cristofoli discusses the network modeling problem when following a network analysis and highlights the fact that the same historical documents can be modeled in different ways [53], which can result in mismatches between the network shape and the research questions. Dufournaud presents her quantitative and network workflow when studying the economic role of women during the 16th and 17th centuries in the city of Nantes, which she splits into three steps: data collection, data processing, and data analysis [67].

3.3 Historical Social Network Analysis Workflow

From the literature and our own projects of HSNA we conducted during the last three years in collaborations with social historians, I propose a formalization of the HSNA workflow divided into 5 steps: *textual sources acquisition*, *digitization*, *annotation*, *network creation*, and finally, *visualization and analysis*. I start by describing the sources and research questions of the different collaborations in §3.3.1, then explain each step of the workflow in §3.3.2, and characterize three properties VA systems supporting this workflow should satisfy in §3.3.3.

3.3.1 Examples

We discussed with four experienced social historians collaborators at different steps of their HSNA workflow about their process: how they inspect and annotate their sources, what network representation they plan to use, and what are their research questions. They all work on semi-structured historical documents, mentioning complex relationships. I provide more details in the following:

1. Analysis of the social dynamics from **construction contracts in Italy in the 18th century [55, 173].** The corpus is made of contracts for different types of constructions in the Piedmont area in Italy. People are typically mentioned under three different construction roles: *Associates* who are in charge of the construction, *Guarantors* who bring financial guarantees, and *Approvers*, who vouch for the guarantors. Documents contain information about the building sites, the types and materials of constructions, and the origins of people. Historians working on this project were interested in characterizing the social structure underlying those contracts, if there were specializations in types of construction, and describing the life trajectory of certain people.
2. Analysis of migrations from the **genealogy of a French family between the 17th–20th centuries** [unpublished work]. The corpus is made of family trees referring to several document/event types: birth and death certificates, marriage acts, military records, and census reports. The social historian wants to characterize the main migrations of individuals and families in France, according to time and place. She is also interested in studying specific families, with theories that, in some areas, people were moving places in a circular fashion over the years. Finally, she is interested in the average social mobility of individuals across the years.
3. Analysis of **marriage acts in Buenos Aires in the 17–19th centuries [169, 201].** The corpus is made of summaries of marriage records that mention the spouses and the witnesses of the wedding. The origin, date of birth, and parents' names are specified for both spouses. The historian is mainly interested in characterizing the relationships between witnesses and spouses—if they are typically from the same family, and if being a witness is sometimes used to ask favors in exchange.
4. Socio-political analysis of **Germans ethnic migration from communist Romania to West Germany in the 20th century (ongoing work) [65].** The corpus is made of administrative forms that mention persons requesting to migrate, along with the persons

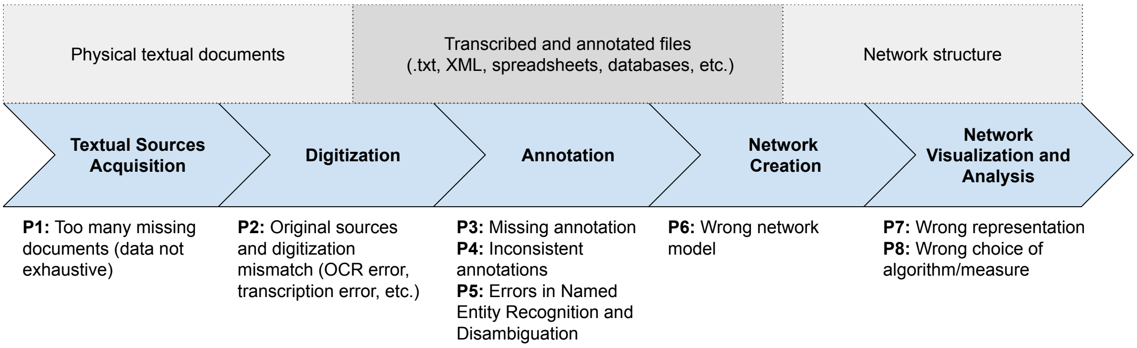


Figure 3.1 – HSNA workflow is split into five steps: textual sources acquisition, digitization, annotation, network creation, and network visualization/analysis. Practitioners typically have to do back and forth during the process. I list potential pitfalls for each step.

they want to join, and the administrative persons of the ministry in charge of the forms. The family members of the aspiring migrants are also mentioned in the forms, with their respective dates of birth. Our historian collaborator is interested in characterizing the socio-economical profile of migrants and the types of family members they are typically joining in Germany.

Each historian planned to follow a network analysis. They typically first read and inspect their sources in depth, before encoding their content with the aim of constructing a network. They plan to use analytical and visualization tools to then explore the structure of the relationships, and answer their questions.

3.3.2 Workflow

I formalize the HSNA workflow of social historians from our collaborations (§3.3.1) but also the literature, and informal discussions with other social historians. We can divide it into 5 steps: *textual sources acquisition*, *digitization*, *annotation*, *network creation*, and finally, *visualization and analysis*. For each step, I present recurring pitfalls which occurred during our collaborations, or that are discussed in the literature [53, 64, 145]. A diagram of the workflow is presented in Figure 3.1.

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

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 tremendously ease in 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. However, most historical primary sources originated before mid 20th century 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 (often called *encoding*) is the process of finding and extracting useful information from the documents concerning the persons, their social ties, and any useful information for the historian. This extra information can concern the persons (their age, profession, sex, ethnicity, etc.) and their social relationships (type, date, place). It encompasses *named-entity recognition* (NER) as well as their resolution. Historians also sometimes annotate information on other entities mentioned in the documents, such as art objects or administrative entities. Usually, historians have a first idea of what they want to annotate in the data as they already explored the documents beforehand and have knowledge of their subject of study, with hypotheses they want to explore. It is, however, common they change their mind through the annotation process, by reflecting on what they found in the documents. Unfortunately, this can produce **missing annotations (P3)** and **inconsistent annotations (P4)** at the end of the process if annotators are not careful. This task can also be challenging, and the choice of annotations has an impact on the final network. Historians also face ambiguity in the process, as several persons and entities (like cities) can have the same name (homonyms), refer to a place name that has disappeared (street name or city), or to an ambiguous person (e.g., John Doe). They, therefore, have to follow a NER and resolution/disambiguation process to identify entities in the sources and disambiguate them across several documents. Entity resolution has always been a problem in social history—as it is more generally in text analysis, where typical groundwork consists in crossing information about the same entities from different heterogeneous sources. However, errors in the disambiguation process can lead to important distortions in the final network structure and properties [64], 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 do doing it manually given the lack of global information (P5)**.

The Text Encoding Initiative (TEI) [50] 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 [68, 209]. 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 one or multiple networks from the annotations of the documents. Typically, all persons mentioned are annotated and are transformed into network nodes (vertices). Additional information, such as their age, profession, and gender, can be stored as node attributes. How the network's links are created is not as trivial and can vary from project to project [6]. The most straightforward approach is to create a link between every pair of persons mentioned in one document, thus forming a clique motif. This is a simplistic heuristic as social relationships can be quite complex, involving more than two persons who can have different roles in the relationship. The choice of the network model has a major impact on the future analysis and **may add bias if chosen loosely (P6)**, such as the creation of network structural artifacts when using network projections [53]. More complex models have been proposed in the literature, such as weighted, dynamic, bipartite, and layered networks, but can be hard to manipulate and visualize. I discuss them more in detail in §3.4.

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 [85, 242]. Usually, historians start to visualize their network to visually confirm information they know and to potentially gain new insight with exploration. Representations need to be chosen wisely given the network as lots of techniques and tools exist for social network visualization. **Some insight may be seen only with some specific visualization technique (P7)**. To test or create a new hypothesis, historians typically rely on algorithms and network measures. Lots of network measures have been developed, like modularity, centrality, and clustering coefficient, that social scientists can leverage to make conclusions [207]. Similarly, social scientists can use data mining algorithms to highlight interesting and potentially hidden structures in the network, e.g., by using clustering algorithms revealing group structures [37]. **However, they have to interpret the results carefully (P8)** as some algorithms act as black boxes and some measures are hard to interpret, with unclear sociological meaning (e.g., centrality). Typically, particular patterns and measure values in the network could have different potential sociological meanings. If we take as an example betweenness centrality which measures the number of times a node appears in the shortest path of every pair of existing nodes, individuals with high values usually highlight positions of power as they communicate with different groups. However, it can also be interpreted as a position of vulnerability in other contexts, such as during periods of wars and repressions, as in the study of Polish social movements in the 20th century by Osa [178], where she shows persons with high betweenness centrality values are more targeted for repression in certain periods. Social scientists, therefore, have to be careful when interpreting network measures and take into account the globality of their sources when interpreting the network they constructed.

3.3.3 Visual Analytics Supported Historical Social Network Analysis

Social historians typically follow the workflow described in §3.3.2 linearly, meaning that at the end of the process, they can realize that the analysis and visualization of the network do not allow them to answer their research questions [146]. This can, in part, be explained by the fact that visualization and analytical SNA tools are only focused on the last part of the process. To fully support social historians, VA interfaces should therefore provide assistance and guidance on the whole process, from the acquisition of the documents (since archives now provide digital catalogs) to the final analysis. Specifically, from discussions with our collaborators, we identify three properties that VA interfaces should satisfy for good integration into the historians' workflow and to limit the recurring pitfalls we identified in §3.3.2: *traceability*, *document reality*, and *simplicity*. First, Traceable systems enable to do easier back and forth between

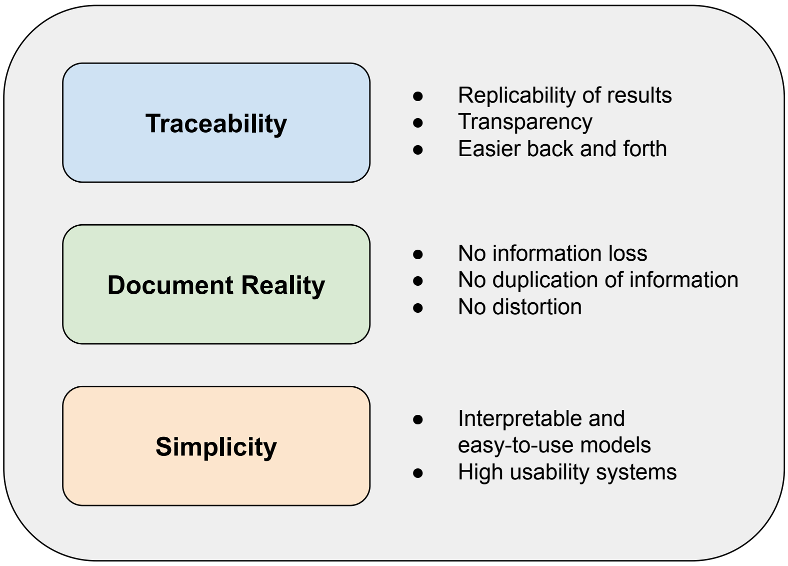


Figure 3.2 – Three properties essential to VA systems supporting the social historians workflow: *traceability*, *document reality*, and *simplicity*.

the different annotation, modification, modeling, and analysis steps and provide a transparent chain of operations leading from the acquisition of the sources to the high-level socio-historical conclusions. Traceability should be operated during the annotation and modeling process (for example to see why two mentions of persons have been given the same identifier, and to trace back network entities to the documents' annotations) but also during exploration. Seeing every low-level operation (filter, selection, group-by, etc.) leading to the generation of insight leads to better transparency and replication [42, 250]. Second, the digitization, encoding, modeling, and analysis/visualization steps should always reflect the textual reality of the documents i.e., the *document reality*[1], in order to reduce the introduction of bias, simplification, and anachronisms in the analysis [125, 147]. Indeed, encoding and modeling the data with abstraction

[1]We chose the term "document reality" over simply "reality" after a conversation with a historian to highlight the fact the historical documents do not describe factually the reality and reflect the subjective bias of the context in which the person wrote them [125]. The content of the documents, therefore, has to be modeled by taking into account this context, which can reveal interesting behaviors and structural patterns. See [148] for specific examples.

and constructed concepts[2] such as the concept of families or "social proximity", often result in distortions (simplification or modifications), duplication, and loss of information contained in the documents. Specifically, the choice of the network model embodies how the content of the sources is manipulated and abstracted with the goal of making historical conclusions, and deeply influences the annotation/encoding and analysis/visualization steps. I discuss network models more in depth in the next §3.4. Finally, as discussed in §1.3, social scientists often have trouble importing their data in SNA tools [6] and often perform "soft SNA" [196] only due to usability problems and "Math anxiety" [181]. VA tools should therefore focus on *simplicity* through the use of simple and comprehensible models and high usability systems. The three properties and their effects on the workflow are summarized in Figure 3.2.

3.4 Network Modeling and Analysis

Historians typically construct one or several networks from their annotated documents that they 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 [67]. 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, I believe that more VA tools should support social scientists in annotating and modeling their documents to make the HSNA process less linear by allowing easier back and forth between the annotation, modeling, and visualization steps. Network models satisfying *traceability*, *document reality*, and *simplicity* properties would mitigate those problems by allowing to navigate more easily between the network and the documents while still modeling well the social relationships mentioned in the sources and being easy enough to visualize and manipulate for analytical and data modification goals.

The choice of the network model to represent the social relationships mentioned in historical documents deeply influence the annotation and visualization/analysis processes. Many network types have been proposed in the literature. While simple ones—which are widely used—are easy to manipulate, they very often break *traceability*—the network entities are not traceable to direct annotations, and sometimes correspond to constructed concepts—and the *reality* of the documents. On the contrary, complex models are often hard to manipulate and visualize. I present the most widely used network models in the HSNA literature in §3.4.1 and present bipartite multivariate dynamic networks as a model satisfying those three properties in §3.4.2.

[2] In anthropology, the terms *emics* and *etics* refer respectively to intrinsic phenomena related to observation and constructed categories and abstractions [107].

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. I describe here the most used network models in HSNA along with more recent ones:

- **Simple Networks [242]:** 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. Moreover, it does not take into account the diversity of social relationships, as every link is identical.
- **Co-occurrence networks [202]:** 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 can be a useful model to detect community-related patterns, but the constructed notion of "proximity" represented by the links simplifies and hide the diversity of social relationships.
- **Multiplex Unipartite Networks [74]:** Only the persons are represented as nodes, and links model social ties between two persons. Links can have different types representing different types of social relationships. It allows the modeling of more complex social relations where people can have various social ties e.g. as parents, friends, and business relationships. However very often several possible representations for the same data exist as projections are often applied to the original documents to get this type of model.
- **Bipartite (also called 2-mode) Networks [102] :** 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 a person and document nodes. Usually, links are not typed and only encode mentions. More recent analyses in HSNA encode the *roles* of the persons in the documents as link types [55]. This network model is more aligned with the original sources and allows following an analysis through the original documents themselves and not through concepts. It can also be used to represent constructed concepts, like the GEDCOM format which introduces the concept of "family" that ties together a husband, spouse, and children with different link types. The concept of family can have different meanings across time and cultures, meaning that GEDCOM adds a conceptual layer instead of grounding the network to concrete traceable documents and events (e.g., no marriage but birth certificates).
- **Multilayer Networks [158]:** 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 [56] and historians [236], 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 [113]:** 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*. Knowledge Graphs are popular for encoding knowledge on the web, including historical knowledge. However, it is

notoriously complex to encode documents using knowledge graphs due to the complexity of the format and the wide choice of possible ontologies. Most historians are unable to understand knowledge graphs and even less to use them for annotating a corpus. Since knowledge graphs are generic, they need complex transformations to be visualized, with no generic system to support historians in taming their high complexity.

Currently, most digital historical projects use unipartite networks (simple, co-occurrence, and multiplex) that are simple and allow answering specific questions, but they do not capture all the complexity of the documents, resulting in simplifications and distortions of the structural patterns. I compare what would be the resulting networks for these models and the bipartite model of our three collaboration use cases (the example #4 is still in the phase of data acquisition), with additional information from the documents encoded as node and link attributes. I do this for one given document for each dataset. The results are shown in Table 3.1.

As shown by Cristofoli [53], we can clearly see the co-occurrence model removes the complexity of the social relationships and only show an abstract "proximity" between individuals. Unipartite multiplex networks allow producing meaningful networks which model well the diversity of relations that can link several people. It especially models simple relationships well, such as parenting ones as in example #2. However, it produces distortions for more complex relationships involving more than two persons, as in example #1 where people can either be mentioned as associates, guarantors, and approbators in the documents. Associates should probably be linked together with *associate* links, but the *guarantors* and *approbators* relationships are more complex to model. Approbators could be linked to the associates, the guarantors, or both. The three ways of modeling this type of relationship make sense but can lead to very different network shapes and analysis results. Historians thus have to decide on a transformation among several possibilities, which will probably distort the social reality of the relationships.

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

Both co-occurrence and unipartite multiplex models thus do not satisfy the *document reality* property by introducing constructed concepts (notion of "proximity") or inferring one-to-one social relationships from mentions in a document mentioning more than two actors.

Moreover, projections add ambiguity in retrospect of the original documents, as it becomes impossible to trace back one link to one specific document, as the same link could potentially refer to several ones [53], i.e., they do not satisfy the *traceability* principle.

More complex models, such as multilayer networks and knowledge graphs, could satisfy *document reality* and *traceability* principles (depending on the modeling choices, as these models are very expressive and do not enforce specific data schemas) but are complex to manipulate and visualize, especially for social scientists. In contrast, the bipartite model satisfies

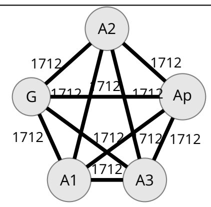
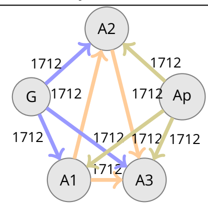
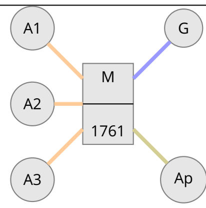
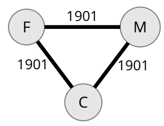
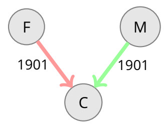
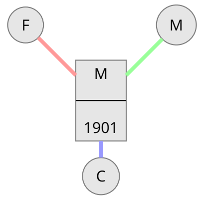
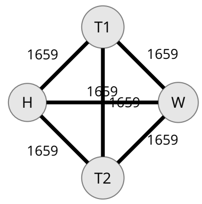
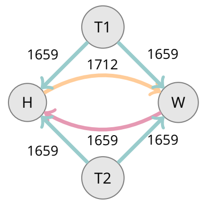
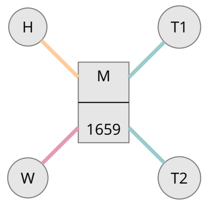
Original Document	Co-occurrence	Unipartite Multiplex	Bipartite
1712: Construction of a church in Torino. Associates: Bellotto G, Bello P.M, Bello G. Guarantor: Astrano G.A. Approbator: Corte A. Associate Guarantor Approbator			
Du dix-neuf fevrier mil huit cent quatre-vingt quatre, à six heures du soir.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

Table 3.1 – Resulting networks using different models produced by one document of the examples detailed in §3.3.1: co-occurrence, unipartite and bipartite models. The first column shows the partial transcription of real documents (simplification for collaboration #1). Colors represent annotations concerning the persons mentioned, their roles, and their attributes. Underlines refer to information related to the events and which can be encoded as document/event attributes. Only time is represented for simplification, but other attributes would follow the same schema. H: Husband, W: wife, T: Witness, M: Marriage, A_N: Associate, G: Guarantor, Ap: Approbator, C: Construction, F: Father, M: Mother, C: Child.

the *document reality* and *traceability* properties through the representation of documents as nodes, and individuals' mentions as links encoding their roles. This model is simple enough to manipulate according to the number of SNA studies leveraging it [61, 152, 176, 211] and the development of SNA bipartite measures and algorithms [30, 102, 142]. Yet, most HSNA studies are based on the network topology and often do not leverage attributes, including time and location. We, therefore, claim that bipartite multivariate dynamic networks allow to model historical documents with *traceability*, *document reality*, and *simplicity* properties. I formalize and describe this model in the next §3.4.2.

3.4.2 Bipartite Multivariate Dynamic Social Network

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The model has, therefore, the following properties:

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

Links and Roles: A link models the mention of a person in a document. **Each link has a type corresponding to the role of the person in the document**. For a marriage act, the roles include *wife*, *husband*, and *witness*. This is a key aspect of our model since it clarifies the relationship between the persons within an event. In contrast, Jigsaw [220] does not consider the roles.

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

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

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

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

Visualization tools using this model can focus on the topology of the network, and/or the attributes which I express here in the format of tuples, commonly used by databases and visualization systems [221]. However, it has to be taken into account that if the attributes extracted from the historical documents are related to vertices and edges independently to the topology of the network, it can be appropriate to compute vertices and edges measures—such as the centrality—and store them similarly to the other attributes, especially so that visualization systems can leverage the same interactions for both. In that case, these types of attributes are directly dependent on potential topology changes in the network (for example after subgraph extraction or network modification interactions).

3.5 Applications

Several tools have been designed for visualizing dynamic bipartite networks that can also be considered dynamic hypergraphs [182, 234], but few incorporate attributes. Moreover, the vast majority of visual analytics tools are solely focused on the analytical part of the data, meaning that the link between the original documents and the hypergraph abstraction is often broken. Social scientists, therefore, always have to do many back and forth between the visual analytics tools and their original documents and the annotation/modeling processes. More visual analytical tools should thus incorporate the textual documents in their data model similarly to Jigsaw [220], as it would allow tracing the entities of the network back to the original documents more easily. Mechanisms to modify the annotations and reflects on the network modeling process directly in the analytical environment could also ease the social scientists' workflow loop. It would allow them to directly correct errors and inconsistencies in the annotations and propagate them in the visual analysis workflow. I propose in chapter 4 and chapter 5 two proof-of-concept interfaces leveraging bipartite multivariate dynamic networks as a representation of social historians sources with the aim of analysis, network modeling, and reflection on the encoding process, with a focus on *traceability*, *document reality*, and *simplicity*.

3.6 Discussion

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

3.7 Conclusion

HSNA is a complex process that starts by collecting historical documents and ends with elaborating high-level sociological conclusions. Historians support their conclusions by modeling individuals' social relationships extracted from the documents and analyzing them through network visualization and analysis methods. Most historical work do not provide details on how they constructed their final network, even though it is a complicated and tedious process that can result in many biases and distortions if not done carefully [6]. We shed light on this process by dividing it into 5 steps and describing recurrent pitfalls we encountered in our projects and collaborations. More importantly, I explain why this process should be done following the principles of *traceability*, *document reality*, and *simplicity* to avoid biasing the analysis, allowing to

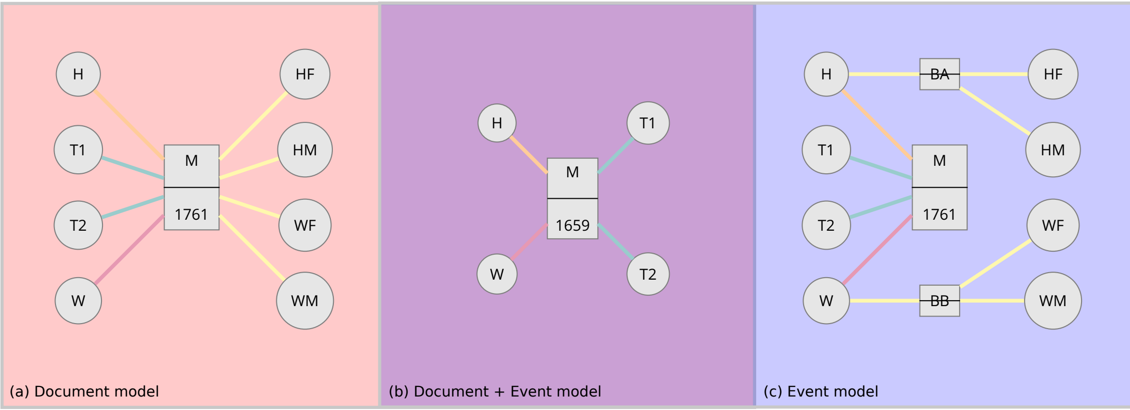


Figure 3.3 – bipartite multivariate dynamic network modeling for two cases of marriage acts of example #3. Some marriage acts mention the parents of the spouses, which is a relationship different than the marriage in itself. This case can be modeled using a document model (a) or an event model (c) by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

go back to the original source at any point of the workflow for easier corrections and replicability, and using models and methods simple and powerful enough for social scientists. Visual analytics software designed for HSNA should consider those principles to provide tools allowing to follow non-biased and reproducible analysis starting from the raw documents while supporting historians in going back and forth more easily between the annotation and analysis/visualization steps. I discussed the network modeling process in depth and claimed that bipartite multivariate dynamic networks satisfy those three core principles, letting historians both wrangle their data and characterize sociological phenomena using a common model and visual representation, thus answering **Q1**. Using this model, VA interfaces could help social scientists manage and analyze their data, starting at the data acquisition and annotations steps instead of focusing on the analysis only while providing efficient representations of the data for analysis and exploration. We explore what could be such VA interfaces in the next two chapters.

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

Contents

4.1 Context . . . **52**

4.2 Related Work . . . **54**

4.2.1 Graphlet Analysis . . . 54

4.2.2 Visual Graph Querying . . . 55

4.2.3 Visual Graph Comparison . . . 55

4.2.4 Provenance . . . 56

4.3 Task Analysis and Design Process . . . **56**

4.3.1 Scenarios . . . 56

4.3.2 Tasks Analysis . . . 58

4.4 ComBiNet System . . . **59**

4.4.1 Visualizations . . . 59

4.4.2 Query Panel . . . 62

4.4.3 Comparison . . . 69

4.4.4 Implementation . . . 71

4.5 Use Cases . . . **72**

4.5.1 Construction Sites in Piedmont (#1) . . . 72

4.5.2 French Genealogy (#2) . . . 73

4.5.3 Marriage Acts in Buenos Aires (#3) . . . 76

4.5.4 Sociology Theses in France . . . 76

4.6 Formative Usability Study . . . **78**

4.6.1 Feedback . . . 79

4.7 Discussion . . . **80**

4.8 Conclusion and Future Work . . . **81**

In the previous chapter chapter 3, I showed that bipartite multivariate dynamic networks constitute a good modeling of historical documents in regard of *traceability*, *document reality*, and *simplicity* properties. However, no visual tools currently exist to specifically explore and manipulate this type of data. In this chapter, I propose a VA interface aimed at exploring historical documents modeled as bipartite multivariate dynamic networks, for historians to be able to reflect on their data and encoding and to follow in depth-analysis. I answer **Q2** by analyzing tasks and questions historians have on their data and providing interaction mechanisms that would allow them to apply specific views and extraction, find errors and reflect on the annotations, and answer their high-level historical question through visual queries and comparison.

This chapter is an extended version of an article accepted with minor revisions to the journal Computer Graphics Forum (CGF) and a poster presented at the conference EuroVis 2022 [188]. It was a joint work with my advisors Christophe Prieur and Jean-Daniel Fekete. I developed the interface and led the discussions, evaluation, and writing of the paper.

4.1 Context

Social scientists such as historians aim to characterize the structure and dynamics of social groups of interest, in a region and period of time they focus on [228]. Their work essentially relies on documents—such as marriage acts, census records, surveys, and business contracts—to gather information about the life of important actors that they explore in-depth, or to draw conclusions on social aspects of groups in the society of that period and place. Instead of drawing conclusions from their gathered knowledge and interpretations of the documents, a more systematic approach consists in constructing a social network from the documents and following a network analysis approach [242]. For this, they need to encode their documents to extract the persons and any other useful information in the text and transfer it into a structured file or a database. Social scientists can then explore, validate, or refute their hypotheses by visualizing and analyzing the network structure and the connectivity patterns between the entities of the resulting network.

Currently, social scientists often model their datasets as simple networks where the nodes are the persons mentioned in the documents (see chapter 3). Usually, Two persons are then connected together in the network when they appear in shared documents. This representation is easy to visualize and analyze but simplifies and distorts the information by hiding the documents that witness the relationships between the persons. Thus, another approach consists in modeling the data as bipartite networks, where both the documents and the persons are represented as nodes and are connected together when a document mentions a given person [96, 199, 211].

In addition, historical documents include time and geospatial information corresponding to the date and location of the events they refer to, and potentially additional information on the mentioned individuals, such as their gender, profession, and date of birth. These are often essential to understanding underlying social phenomena, as time, space, and classical social categories play an important role in sociological structures and dynamics [146]. For these reasons, as we discussed in chapter 3, historical sources and the underlying social events they refer to can be modeled well by *bipartite* with *roles*, *multivariate dynamic* networks. *Bipartite* means that both persons and documents (or events, that are often witnessed by physical documents) are modeled as typed nodes. *Multivariate* means that the nodes and links can carry additional attributes. *Dynamic* means that time is a mandatory attribute of documents. Furthermore, a link created between a person's node and a document's node (when the person is mentioned in the document), has an associated link type that models the *role* of the person in the document/event. Additionally, documents can optionally carry a geographical location. This model unifies several social network models and allows the representation of historical documents with *simplicity*, *traceability*, and *document reality*, i.e., the relationships appear as they are mentioned in the documents without distortions implied by projections [53].

More complicated models exist, such as knowledge graphs [177], which are very expressive but hard to manipulate, especially for social scientists. In contrast, most visual and analytics tools widely used by social scientists such as Gephi [18] and NodeXL [218] enforce too simplistic network models and only provide limited interactions for exploring the network data, even if they provide the computation of several network measures. This results in many social historians ending their analysis by plotting the network using a node-link diagram—which is hard to read with dense networks—, and identifying the most important actors with the help of centrality measures [147]. Lemercier et al. describe this phenomenon the following: "Network graphs of the 'spaghetti monster' variety are a case in point. Often, in historical papers, they are used to show that a network is dense (and that the author has mastered the new technology). The narrative then comments on the individuals identified as central in the network. This approach can indeed be quite interesting, but it is hardly the only possible use for network analysis. [147]". VA tools guiding social scientists towards more complex exploration with the help of interaction mechanisms, high usability and interpretability are therefore needed.

In this chapter, I present a VA system to explore and analyze Bipartite Multivariate Dynamic Social Networks, with the aim of supporting more complex historical analysis based on easy-to-use interactions, but also potential data correction. I elaborated the tool based on four collaborations with social scientist colleagues. I first collected important questions they each had on their data and transcribed them from a network analysis perspective. The majority of the questions raised consisted in either finding specific patterns in the network—corresponding to specific groups or individuals exhibiting intriguing behaviors—or in comparing several subsets of the network, in terms of network measures, attribute distributions, and their overlaps. I hence focus on three high-level tasks: exploration, queries, and comparison of this type of network. Users can explore the data using two layouts: a node-link bipartite view showing the sociological structure of the network, and a map layout based on the geolocation of documents. I designed and implemented a new visual graph query system that allows us to build both topological and attribute constraints, based respectively on a node-link interactive representation, and dynamic widgets. By easy-to-create queries, social historians are able to 1) detect erroneous patterns and reflect on their encoding process and 2) find relevant patterns which can answer their historical questions. For this, I rely on the *Neo4j* graph database [162] and its query language, *Cypher*. Most visualization systems offer dynamic queries to hide the complexity of query languages.

However, using a rich data model, some queries are easier to refine using scripting than dynamic queries. I implemented dynamic queries that also show the translated Cypher queries, and inversely, can translate textual queries into visual queries. With that interface, social scientists can start building their queries with simple widgets and, if needed, complement them by editing the query, alone or with the help of power users. Furthermore, they can export their query, the associated results, and its history at any point to share it with someone else or to start an analysis session from a previous result. ComBiNet also implements subgraph comparison techniques, allowing the comparison of networks, network-related measures, and attribute distributions between the entities returned by the queries. I validate ComBiNet with the description of four real-world use case showing the system can be used to answer socio-historical questions while reflecting on the annotation/encoding process, and a formative usability study demonstrating the system can be used smoothly by social historians.

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

4.2 Related Work

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

4.2.1 Graphlet Analysis

One of the inspirations for this project came after participating in the 2020 VAST challenge[1] where our team used graphlets to measure the similarity between several networks [230].

Graphlets are small connected induced, non-isomorphic subgraphs composing any network (see §2.3.2 for more details). They were first introduced by Milo et al. [164] to explore the structural differences between biological networks, but they are now used in several disciplines involving networks such as sociology [46].

One of the aims of the VAST 2020 challenge was to compare several multivariate networks. However, by using graphlets, we realized that 1) it was not very efficient to compare several networks in contrast to other measures, and 2) the interpretation of all graphlets patterns that are found in a network is complicated given the fact that one specific pattern can have various interpretations given the nodes involved and their positions in the network [118]. This is especially true that the number of potential graphlets grows exponentially as we increase

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

the number of nodes considered (there are 6 graphlets of size 4 and 21 graphlets of size 5 for example) and if we add complexity to the network model, for example by using directed links or node and link types [194].

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

4.2.2 Visual Graph Querying

Graph pattern matching is an important task in SNA, which consists in finding a subgraph of interest in a larger graph [76]. Several scripting languages, such as *R* [225] and *Python* [235] have been extended to support the exploration of social networks using specialized libraries such as *igraph* [58] and *NetworkX* [101], and provide functions for graph pattern matching. However, social scientists are often challenged to use scripting languages and programming, as they often do not have formal training in such technologies. Graph databases allow the storage and manipulation of network data with the use of query languages, such as the Cypher language for Neo4j [162]. To lower the complexity barrier of their usage, several visual graph query systems have been developed, enabling analysts to rapidly build and refine their queries visually. Some systems hide the scripting language, such as *GRAPHITE* [47], *Intuinet*[2], and *VERTIGo* [59] that allow specifying a graph query as a node-link diagram that the user creates interactively. Shadoan and Weaver [210] use a similar concept with hypergraphs to filter multidimensional data. Other systems, such as *VIGOR* [185], only visualize the query after it has been written by the user with a scripting language (here Cypher) and do not allow direct interaction of the visual representation of the query. All these visual query systems are limited to topological queries with constraints on the vertex and edge types and do not allow to make constraints on other dimensions, such as attributes and time associated with vertices and edges.

4.2.3 Visual Graph Comparison

Gleicher et al. [93] propose a taxonomy of visual comparison designs for complex objects. They claim any visual comparison system can be classified into one (or a mix) of the three following categories: juxtaposition, superposition, or explicit design. Yet, few visual systems support comparison tasks for social networks.

Andrews et al. [8] describe a technique to compare several networks, using a combination of juxtaposition and superposition techniques. The two candidate networks are shown side by side, along with a third view composed of a fusion network highlighting both the shared nodes along with the non-shared nodes with different colors.

Freire et al. [86] 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 [105] describe a system to match and compare graphs using superposition, focusing on the topology, not taking into account attributes or time. Tovanich et

[2] See `https://intuinet.fr`

al. [230] propose a visual analytics tool to compare multivariate, sometimes bipartite, dynamic networks and find common structures. However, the tool does not handle roles and is designed for the specific task of matching a subgraph into a larger network.

4.2.4 Provenance

Provenance in the context of VA 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 with action recovery (undo) plus collaborative and reproducibility capabilities [191]. For example, the VisTrails system allows users to reproduce their visual analyses by providing an executable history graph of their actions [42], while GraphTrail provides provenance tools to ease collaborative analysis [70]. Provenance can also be beneficial for visualization designers and researchers, as it gives them a tool to understand users' behaviors [21, 31] and evaluate/improve visualization systems [193]. All the reasons and concrete implementations of provenance are discussed in depth in Xu's survey [250].

4.3 Task Analysis and Design Process

I designed the ComBiNet tool in collaboration with social historians who wanted to follow a network analysis on their historical semi-structured documents, that are well modeled by bipartite multivariate dynamic networks. I first collected questions they had about their data and what they wanted to see in a visual interface. By analyzing the questions, we leveraged tasks and requirements that I used to design and implement the interface, with continuous feedback from our collaborators.

4.3.1 Scenarios

We elaborated this interface from the collaborations with historians I described in §3.3.1. These collaborations involved regular meetings and multiple discussions over three years. All these datasets are textual corpora constituted of historical documents mentioning people with complex relationships, which are well modeled with bipartite multivariate dynamic networks. We give more details about the datasets of these collaborations in this section, along with our collaborators' main questions with the associated network queries to answer them. The full answers involve visualizations of the query results and attribute summaries that I describe in the next section. We categorized the questions according to four dimensions: global (G)/local (L) (do they want to categorize a group of nodes or retrieve specific persons/documents), if the question can be answered using the topology (T), and/or the attributes (A), and finally if a comparison (C) using several filters is needed or not (N).

1. Analysis of the social dynamics from **construction contracts in Italy in the 18th century (141 documents, 272 persons) [55].** The corpus is made of contracts (manuscript documents) for different types of constructions in the Piedmont area in Italy. People are mentioned in three different roles: *Associates*, who participate in the construction; *Guarantors*, who bring financial guarantees; and *Approvers*, who vouch for the guarantors. Along with the time and location of the construction site, documents have a construction type (military, religious, and civil), work type (big work, small work, reparation, transportation, etc.), and material (wood, stone, metal). People also have

an origin attribute (the place they come from), manually extracted from the original documents.

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

Query 1.1 Request all approbators occurrences

Query 1.2 Request all associates and guarantors occurrences

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

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

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

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

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

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

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

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

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

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

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

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

2. Analysis of migrations from the **genealogy of a french family between the 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 What is the trajectory of life for a given specific individual (birth, living, marriage, death) (L, A, N)

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

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

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

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

Query 9.1 Select persons with a geolocated birth and death certificate

Question 10 Are there differences in volume and location between migrations in the 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 19^{th} century

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

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

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

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

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

3. Analysis of migrations from Spain to Argentina through the **marriage acts at Buenos Aires in the 17–19^{th} centuries (1381 acts, 6659 persons) [169].** The corpus is made of acts that mention the spouses and the witnesses of the wedding, which are the roles modeled by the links. The origin, date of birth, and parents' names are specified for both spouses.

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

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

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

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

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

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

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

4. Socio-political analysis of **migration of ethnic Germans from communist Romania to West Germany in the 20th century (ongoing work) [65].** 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 members of their family do emigrants usually join? (G, AT, N)

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

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

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

4.3.2 Tasks Analysis

Most of the questions we collected from our collaborators could be answered by isolating a subgroup of entities and analyzing them in the context of the whole network, or by comparing two subgraphs, in terms of their entities, structure, and attribute distributions. From discussions with our collaborators and the analysis of their questions on their data, we elaborated a list of requirements for the visual interface, split into three main parts: 1) Exploration of the data, 2) Queries, and 3) Comparisons. The elaboration of the tasks was an iterative process, as we showed the interface to our collaborators several times in the development phase to get feedback. The tasks are described here and summarized in Table 4.1:

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

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

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

4.4 ComBiNet System

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

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

4.4.1 Visualizations

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

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 the network's context	V1, V2	
Compare several subgroups	T3.1 Show the shared and exclusive entities	V1/V2	
	T3.2 Compare the node attribute distributions	V4	
	T3.3 Compare the subgraph measures	V3	

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

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

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

V1: Bipartite Node-Link View The bipartite node-link visualization panel shows the network using the DrL force layout from igraph [58] with overlap removal using D3 [32]. Node-link representations are very common in social sciences [18, 54, 171, 218] and were a specific request from our collaborators. In the context of our bipartite model, the persons are represented as circles and the documents/events as squares, while the roles are encoded as link colors. A link models the mention of a person in a document. This view provides an overview of the data by showing the structure of the network (T1.1) and the roles of the persons in their different documents (T1.2). The view also provides pan & zoom and selection interactions for effective navigation. Nodes' labels are displayed (name of person and ids of documents by default) on the canvas with an occlusion-free mechanism that hides nodes with a low degree when two or more nodes' labels overlap.

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

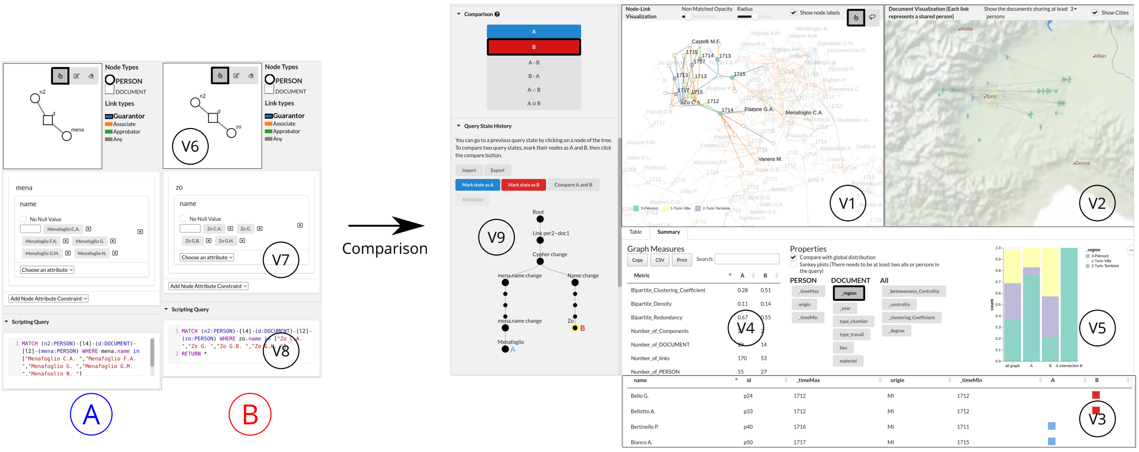


Figure 4.1 – The ComBiNet system was used to compare two subgroups of a social network of contracts from [55], extracted with dynamic visual queries. (A) and (B) show the two visual queries created by the user in the query panel using an interactive node-link diagram editor (V6), dynamic query widgets (V7), and the equivalent Cypher script (V8). The right part shows ComBiNet's global interface in *comparison* mode: (V1) Network visualization panel, (V2) Map of the geolocalized nodes, (V3) Table of persons, (V4) Graph measures comparison, (V5) Attribute distribution plots, and (V9) Provenance tree. The two visual queries on the left, translated into Cypher queries below, select the "Menafoglio" family on the left, and the "Zo" family on the right, along with their construction contracts and close collaborators.

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 [71].

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

V3: Entities Tables All the persons and the documents of the loaded dataset are listed in two separate tables, showing the attributes of the entities (person or document). This way, users can order the entities according to any attribute they want (T1.2). The tables are linked to the visualizations, meaning that selecting a row highlights the respective entity in the visualizations and vice-versa. Selecting a node hence highlights the corresponding row and pushes it to the top of the table. Tables in social network visualization systems have been proven to be efficient and useful for social scientists when exploring their data [26] and are a feature that has been asked by our collaborators. It allows them to link the visualization to the network entities more easily, and dive deeper into one entity's attribute values after selecting it in the network. For example, if the visualization reveals an intriguing person connected to two distant components through two documents, the user can rapidly see the information of this person and documents on the tables, to see if this could be an error from the annotations or an interesting person he or she could investigate more in depth in the original sources. It also makes ranking entities according to various criteria easier and more straightforward. Finally, the tables are exportable in CSV, pdf, or directly in the clipboard, which was a request to our collaborators.

V4: Graph Measures The Graph Measures view shows measures related to the network and gives insights into its structure to users (T1.1). We report simple measures like the number of persons, documents, links, and components, and more sophisticated bipartite network measures asked by our users, that they can report for their analysis: the bipartite density, bipartite average clustering coefficient, and bipartite average redundancy [142]. These measures are updated in real time when filters and comparisons are applied.

V5: Attributes View All the attributes in the network are shown as buttons in the bottom right of the interface, sorted by their associated node type (person, document, and "All" for both types). They can be quickly visualized by hovering over the button, producing two effects: it colors all the nodes on the two views according to their attribute values, and it shows a plot of the distribution of the selected attribute. Figure 4.2 shows the construction dataset of collaboration #1 where the user selected the *_year* attribute, coloring the document nodes with their year in the node-link diagram (left) and the map view (right), revealing for example that most construction occurring in 1714 occurred in Torino and Torino's surroundings. By clicking on the button, the visual encoding and the distribution plot remain selected. This interaction is inspired by the x-ray technique of the Vizster system [109]. 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).

4.4.2 Query Panel

The query panel allows users to rapidly build queries visually, with topological and attribute constraints. The visualization of the query is synchronized with the Cypher query sent to the database. Modifying one representation update the other, allowing users to build a query visually and refine it in Cypher when appropriate. Experts users who know the Cypher language can also start to construct their query textually and modify it visually later on. In this subsection, I describe all the features and interactions allowing ComBiNet to build a query and illustrate them with questions 2 and 6 of our collaboration #1. Our collaborator wants to *find the persons who are mutually guarantors to each other in separate contracts (2)* and to know *how Torino and Torino's surroundings differ according to their contracts (6)*. Figure 4.4 (left) shows the final queries, but first, I explain how to create them. ? **V6: Node-Link Dynamic Query**

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

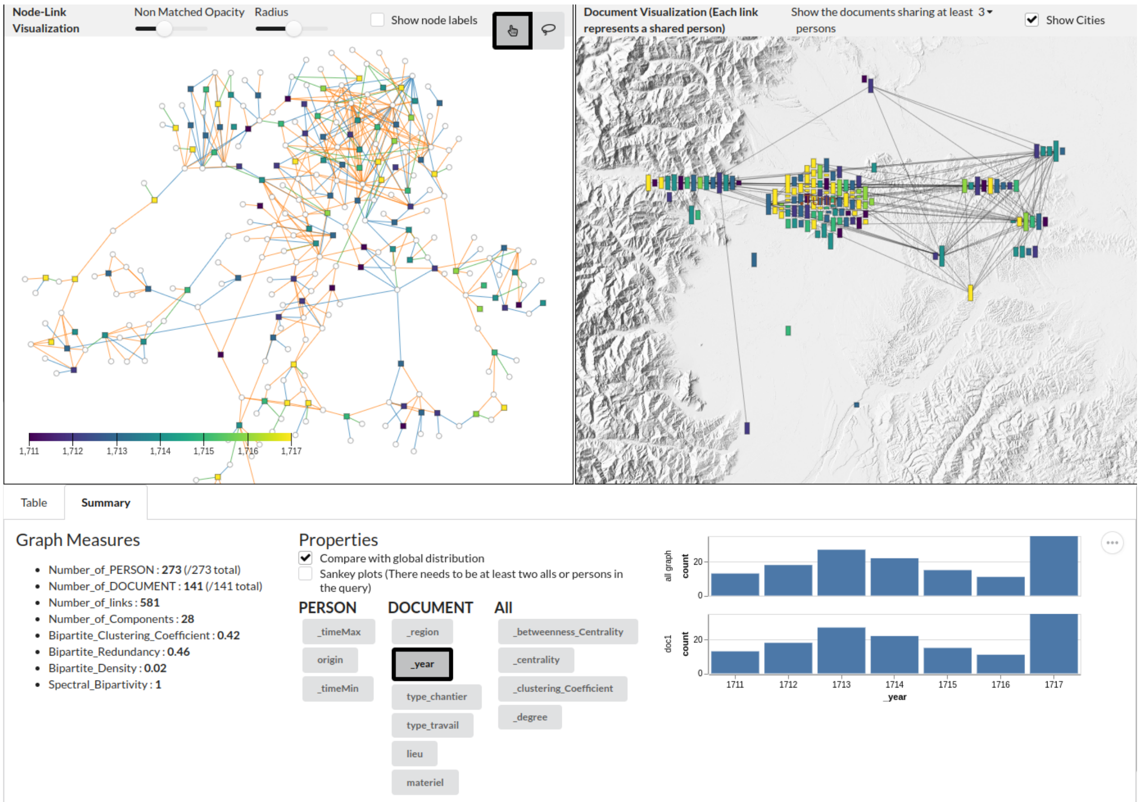


Figure 4.2 – CombiNet interface wreal-timeith the dataset of collaboration #1. The user selected the *_year* attribute, showing the distribution of document years with a histogram (bottom right), and coloring the documents node on the bipartite view (left) and map view simultaneously (right).

highlighted in the main visualization views (V1, V2). Three modes of interaction are available through the top-right menu: *selection*, *addition*, and *deletion*. The *selection* mode allows to drag the nodes in the panel, while the *addition* and *deletion* modes allow the following actions:

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

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

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

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

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

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

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

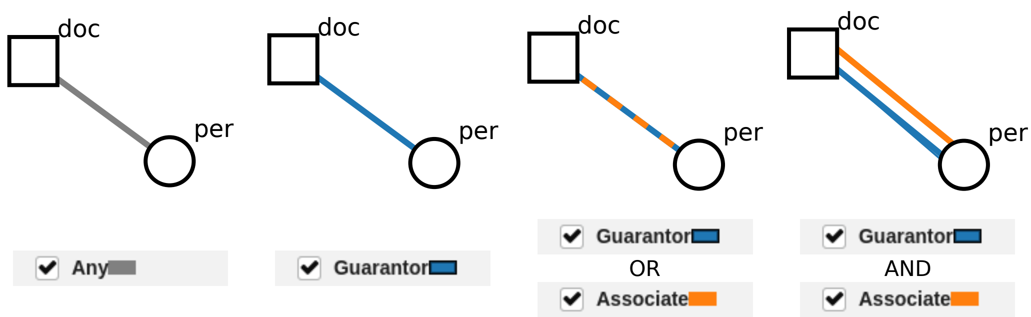
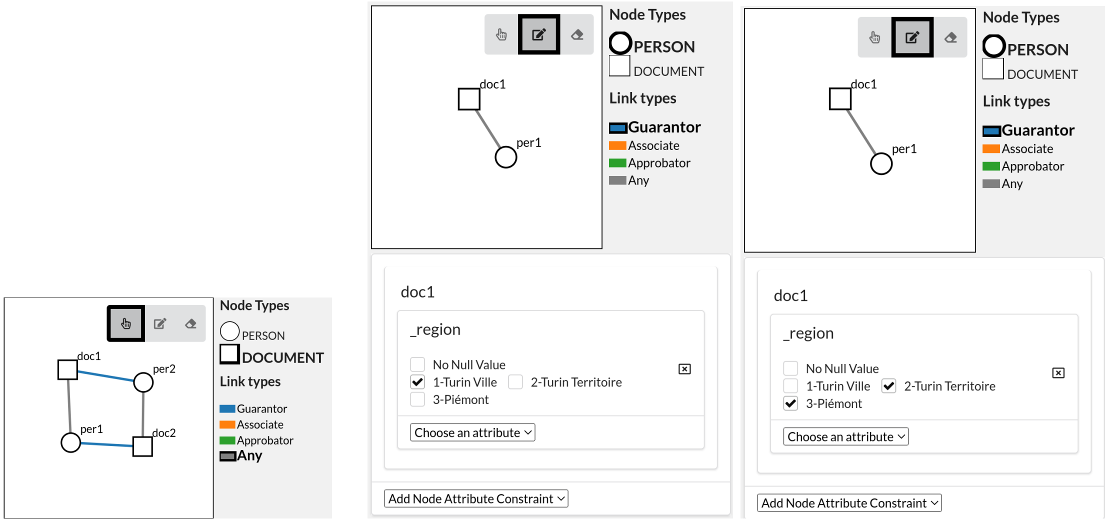


Figure 4.3 – All link creation possibilities: Any link type (left), one selected link type, here guarantor (middle left), the union of several link types (middle right), several links with different types (right)

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

V7: Attribute Constraint Widgets Users can also add attribute constraints (T2.2) on the created nodes with the help of interactive widgets. An input button is created for each node and link identifier from the node-link query panel. It allows the creation of a dynamic query widget for any of its attributes. The widget design varies according to the three possible attribute types: numeric, categorical, and nominal, as in the original dynamic queries formalization by Shneiderman [214]:



(a) (b)

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

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

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

To answer our collaborator's second question (*how do Torino and Torino's surroundings differ according to their contracts?*), we first filter the documents which are located in Torino (*Turin*). For this, we select the whole dataset by linking a person and document node with an untyped link. Then, we select the id *doc1* of the document of our visual node-link query, and the *region* attribute. It initializes a categorical widget including all the values found in the dataset for this attribute with associated checkboxes. We check the region of interest "*1-Turin Ville*" to select all the documents from this region. The first widget of Figure 4.5 illustrates the created constraint. To select the documents of Torino's surroundings, we can simply uncheck the "*1-Turin Ville*" value for the *region* attribute and check the two other values "*2-Turin Territoire*" and "*3-Piemont*" which are areas corresponding to the surroundings of Torino. Both queries are represented in Figure 4.4 (b).

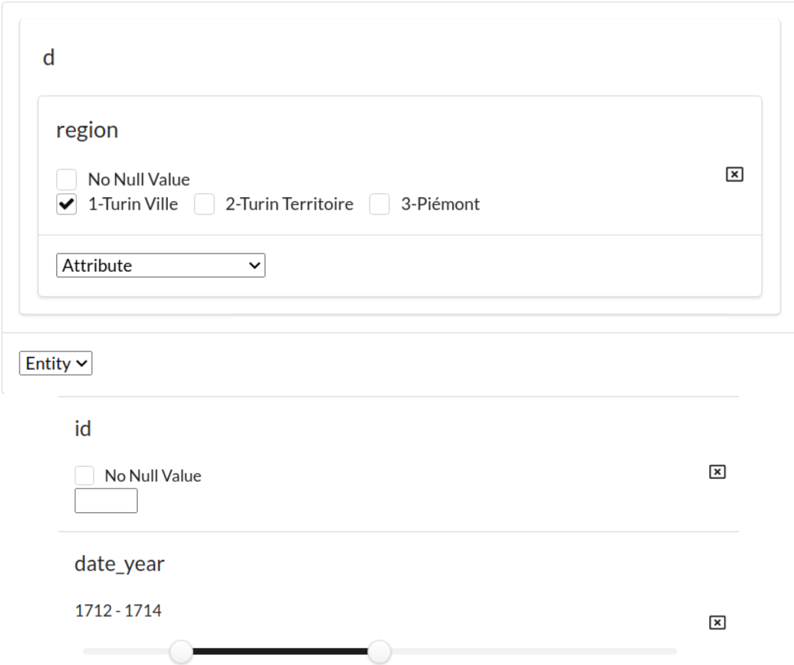


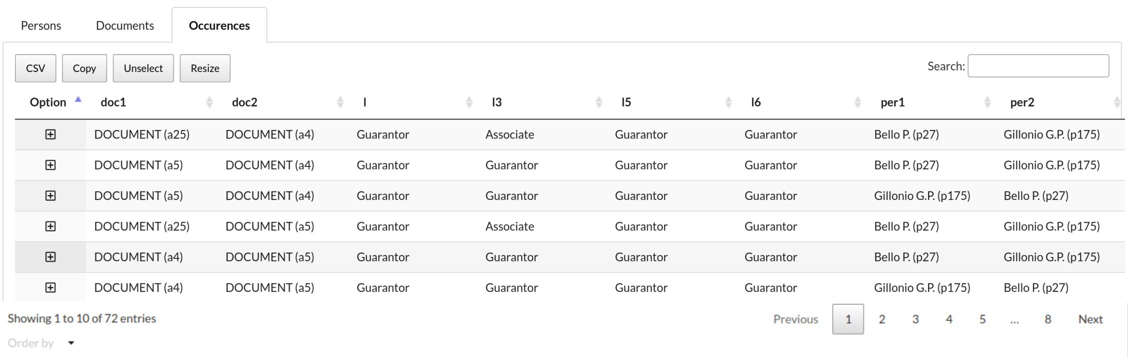
Figure 4.5 – Widget designs for the different attribute types: checkboxes for categorical attributes (top), text input for nominal attributes (middle), and a double slider for numerical attributes (bottom). The categorical attribute example shows the inputs letting users create new constraints for other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

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

Query Results

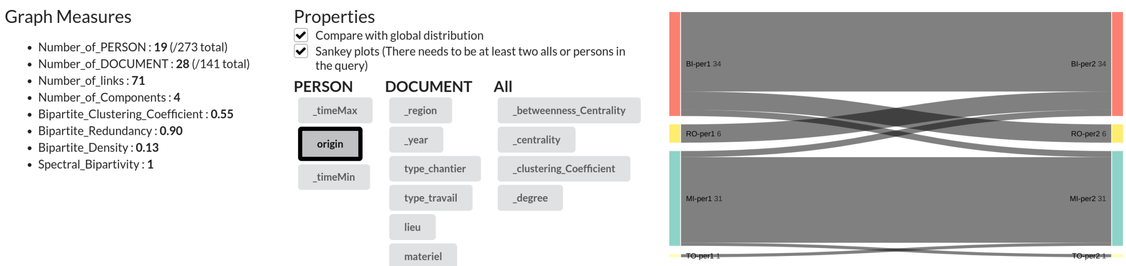
Each modification of the query, whether from the node-link dynamic query, the widgets, or the Cypher text box, update the two visualization panels (V1, V2), the entities tables (V3), the network measures view (V5), and the attribute plots (V6). The nodes and links that do not match (are not retrieved by the query) are grayed out in V1 and V2 and are removed from the persons and documents tables (V3). A third table shows every occurrence found of the created pattern that we call the occurrence table. The occurrence table for question 1 of collaboration #1 is shown in Figure 4.6 (a). It tells us that the occurrence has been found 72 times, meaning that the pattern exists 36 times in the network by taking into account the symmetry of the subgraph query. Users can switch between the three tables in the table view using the tabs.

The network measures are computed on the new subgraph formed by the union of all patterns found and updated on the network measures view (V5). Figure 4.6 (b) (left) shows to the user the different graph measures of the subgraph induced by the patterns found. Since some measures can be long to compute, the values are computed iteratively in the backend and shown progressively [78] 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.



Option	doc1	doc2	l	l3	l5	l6	per1	per2
⊞	DOCUMENT (a25)	DOCUMENT (a4)	Guarantor	Associate	Guarantor	Guarantor	Bello P. (p27)	Gillonio G.P. (p175)
⊞	DOCUMENT (a5)	DOCUMENT (a4)	Guarantor	Guarantor	Guarantor	Guarantor	Bello P. (p27)	Gillonio G.P. (p175)
⊞	DOCUMENT (a5)	DOCUMENT (a4)	Guarantor	Guarantor	Guarantor	Guarantor	Gillonio G.P. (p175)	Bello P. (p27)
⊞	DOCUMENT (a25)	DOCUMENT (a5)	Guarantor	Associate	Guarantor	Guarantor	Bello P. (p27)	Gillonio G.P. (p175)
⊞	DOCUMENT (a4)	DOCUMENT (a5)	Guarantor	Guarantor	Guarantor	Guarantor	Bello P. (p27)	Gillonio G.P. (p175)
⊞	DOCUMENT (a4)	DOCUMENT (a5)	Guarantor	Guarantor	Guarantor	Guarantor	Gillonio G.P. (p175)	Bello P. (p27)

(a)

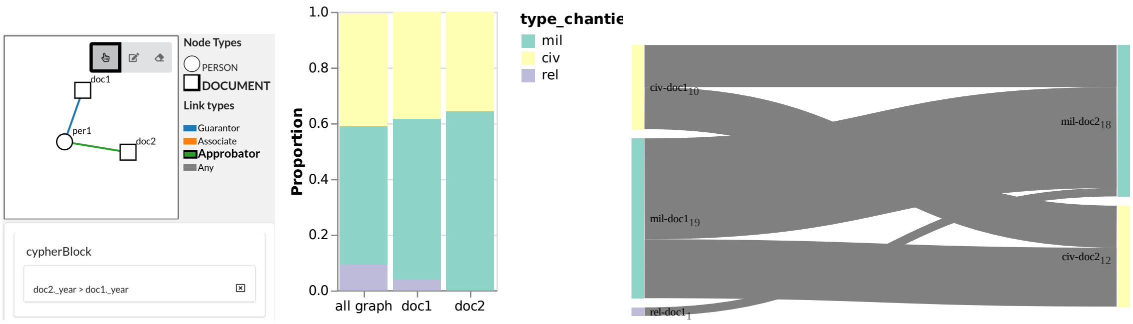


(b)

Figure 4.6 – Results of question 2 of collaboration #1: (a) shows a subset of the table view with every occurrence of the pattern found. (b) shows the summary panel, with the graph measures and the attributes view with the *origin* attribute selected and the Sankey option checked. It allows us to see the attribute distribution of the persons included in the pattern and see if there is a relationship between persons who are mutually guarantors and their origin.

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

The attribute plots are exportable in SVG, while the tables are exportable in CSV, PDF, or directly in the clipboard. This was a demand of our collaborators, so they can export their results easily for another analysis or for communication purposes.



(a) (b) (c)

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

The graph measures and attributes views for the results of question 2 of collaboration #1 are shown in Figure 4.6. The Sankey view of the *origin* attribute shows that mutual guarantors come from 4 regions only and that usually, people have mutual guarantor relationships only with persons of the same origin. This is especially true for persons from *Milano*, and with some reciprocal links between persons from *Bioglio* and the *Comune di Ro*.

V9: Provenance Tree Each change in the query panel is saved with the computed results so that the history of the query construction can be shown in the form of a provenance tree (T2.4), managed with the Trrack library [60]. Each node of the tree represents a query change, with a descriptive label such as "New Link". It enables to rapidly visualize the succession of filters applied with their refinements. At any moment, users can rename a tree node or click on it to go back to the previous state; allowing them to explore different query possibilities easily and iteratively. Hovering over a node shows a tooltip with the query panel associated with the selected query state. It lets 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, permitting to revisit and refine explorations. Figure 4.8 shows the provenance tree made to answer question 2, split into 2 branches, with the tooltip showing one of the node query states. The whole provenance tree is exportable and importable in JSON format, allowing to 1) start

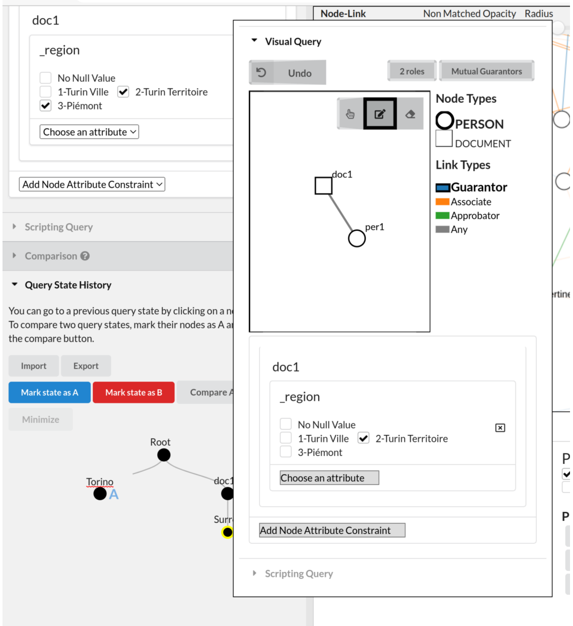


Figure 4.8 – Provenance tree to answer question 2 of collaboration #1: left branch leads to Torino documents (the node is labeled as A) while right branch leads to surrounding documents (the node is labeled as B). The user hovers over one node, revealing a tooltip that shows the visualization of the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

a session from a previous exploration and not from scratch, 2) share exploration sessions and results with others, and 3) provide a trace of the exploration leading to a potentially interesting result, hence providing *traceability* in the results.

4.4.3 Comparison

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

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

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

Topological Comparison In comparison mode, users can rapidly switch between the visual filters of (A) and (B) by hovering over their respective buttons on the comparison menu and thus compare the structure of the two resulting subgraphs (T3.1). Similarly, different boolean comparison operations are available by hovering their respective buttons (Figure 4.1-C), such as the intersection, union, and differences between the two filters. Moreover, the summary tab allows comparing the different graph measures of the two subgraphs by showing them side by side (T3.3). Figure 4.9 shows the comparison table for the queries returning the subgraph of Torino (A) and Torino's surroundings (B). Comparing these measures, such as the number of matched documents or the densities, is crucial for SNA. For example, the table indicates that the density is two times higher for Torino, suggesting that fewer persons participate in the same construction compared to Torino's surroundings.

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

Showing 1 to 8 of 8 entries Previous 1 Next

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

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

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

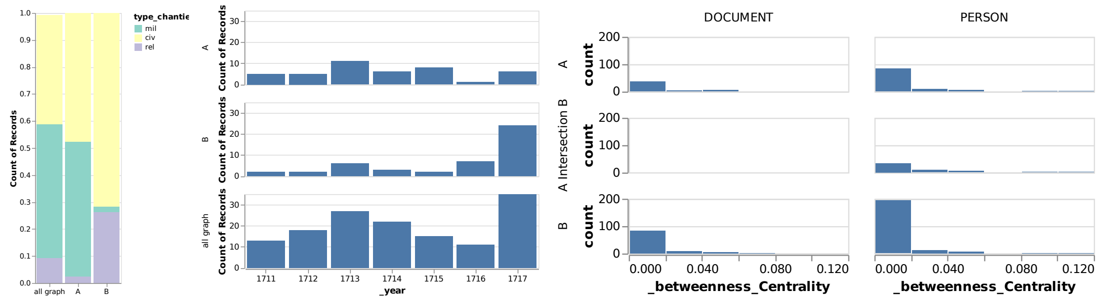


Figure 4.10 – Distribution of the type of constructions, the years, and the betweenness centrality for the documents and signatories of Torino (A), Torino's surroundings (B), and the whole graph (top).

4.4.4 Implementation

ComBiNet[3] is made of three components: a visual web interface, a python server, and a Neo4j [162] graph database instance. The client interface is written in JavaScript using D3 [32], Vega [204], and the Trrack library [60]. The python server is written in Flask and interacts with the Neo4j instance for query processing before sending the results to the frontend. We implemented the Cypher parser with the ANTLR parser generator [180]. The abstract syntax tree of the Cypher query is used as a representation of the query. Modifying the query visually

[3]The web application and source code are available at https://www.aviz.fr/Research/Qcompnet

updates the tree, which is translated into Cypher in the textual query panel. Similarly, a manual change in the Cypher query updates the abstract syntax tree which is translated into a visual query.

4.5 Use Cases

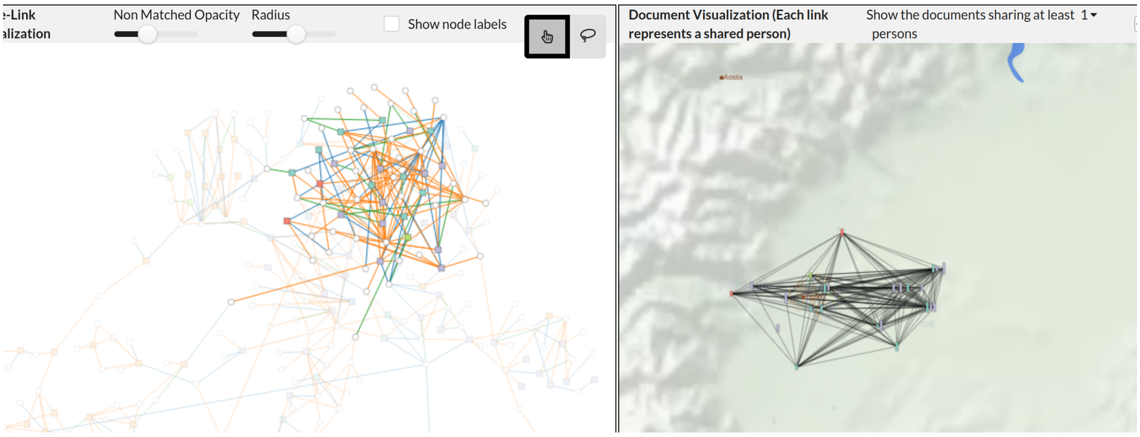
In this section, I describe how our system has been able to specifically answer questions from three of our collaborators and one other use case. The tool was mostly operated by the developers working side by side with the collaborators to test the expressiveness of the queries and the value of the results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

4.5.1 Construction Sites in Piedmont (#1)

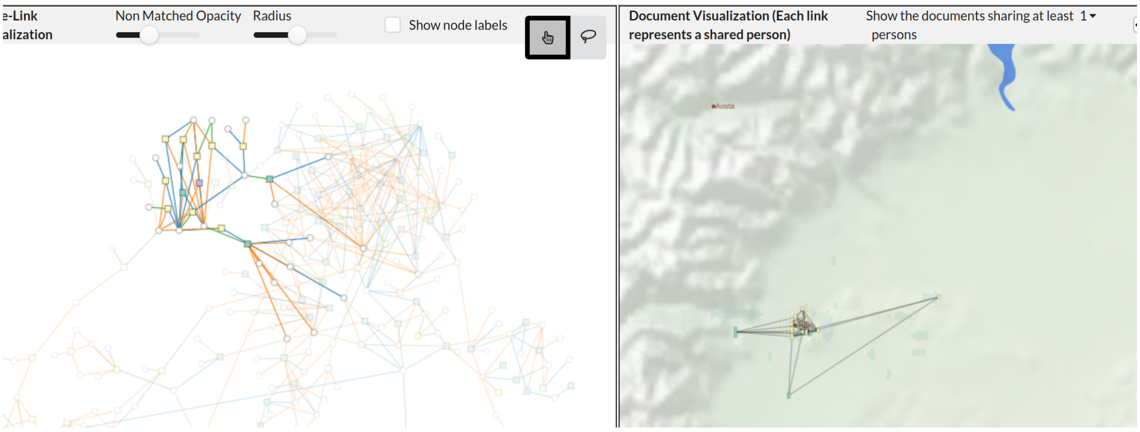
One of the main questions of our collaborator was to compare two families which he knew played a big role in the structure of the network: the *Menafoglio* and *Zo* families (question 4 in Table 4.1). Specifically, he was interested in knowing if there were differences in specialization in the type of 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 4.11 (a), we can see that the group is pretty dense with people collaborating a lot between them. Looking at the map, we can clearly see that the family has been mostly active in Piedmont outside of Torino and Torino's close territory. We also have a first view of the attribute distribution of the persons in the group and their contracts. We then do the same query for the Zo family. We keep the same topological filter and replace the name filters with the core members of the Zo family known by the historian. We see on the graph view (Figure 2 of the supplementary material) that the group is smaller and is in a different area in the graph. The map enriched with a selection of the *region* attribute shows that, contrary to the Menafoglio, the Zo family has been more active in Torino and Torino territory (a very close area of the city). The two groups can be compared using the *comparison mode* by selecting the two queries in the provenance tree. This opens the comparison menu to quickly navigate between the visual selection of (A), (B), and the set $A \cap B$ that interests our collaborator. The table showing the graph measures of the two subsets confirms what is shown visually: the Menafoglio group is more populated but less dense than the Zo family.

Our user is then interested in comparing the distribution of several attributes between the two groups. We can clearly see in Figure 4.12 (top right) that the Menafoglio family is more specialized in military (*mil*) sites, while the Zo family is doing more civil (*civ*) 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 on average (bottom right, middle chart). This makes sense as they act as bridges linking the two families.



(a)



(b)

Figure 4.11 – Menafoglio (a) and Zo (b) families were retrieved with queries and highlighted in the bipartite node-link and map views.

4.5.2 French Genealogy (#2)

We describe how ComBiNet allowed us to answer an important question of the use case #2: to detect the largest migrations across several generations, in which areas, and at what time they occurred (question 9 in Table 4.1). The map view shows at a glance (Figure 4.13) that the majority of events have taken place in three specific regions: west (Britany), mid-north (Paris), and mid-south (Limousin).

To find patterns of migrations within families, we first make a query representing a simple family by linking a person node to a birth event, connected to the parents using a link of *father* or *mother* type. We repeat the process to the new parent node to add another generation. Finally, we connect the latest generation child with a death event, to have another date and location to compare to (see Figure 4.14a). 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

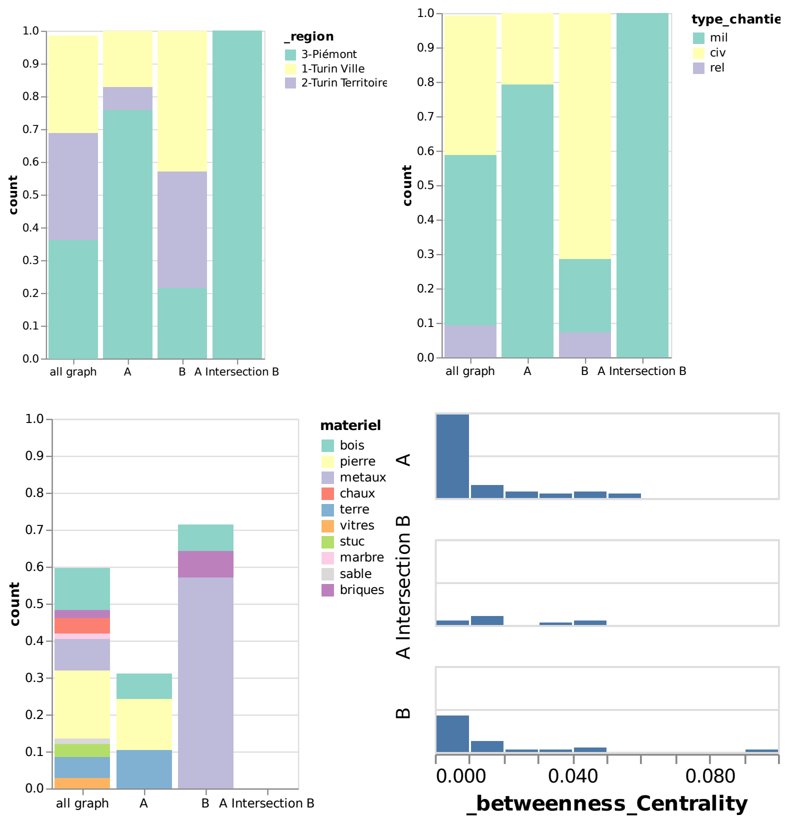


Figure 4.12 – Attributes distributions plots between the whole graph, the *Menafoglio* family (A), the *Zo* family (B), and $A \cap B$, for the *region,type_chantier, material type.*

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 families. Figure 4.14b shows that the majority of families are from *Haute-Vienne* (which can easily be confirmed by checking the map), and do not move much across generations. Our collaborator however detected interesting patterns of people moving from the department *Creuse* to *Haute-Vienne* across two generations. She refined the query by adding an attribute filter on this specific department using a widget. The table view then showed her who these migrants were and when it occurred. The bipartite visualization panel allowed exploring more in-depth this specific group of people.

Afterward, we answered question 9 (Table 4.1), i.e., is there a significant difference in 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 at the

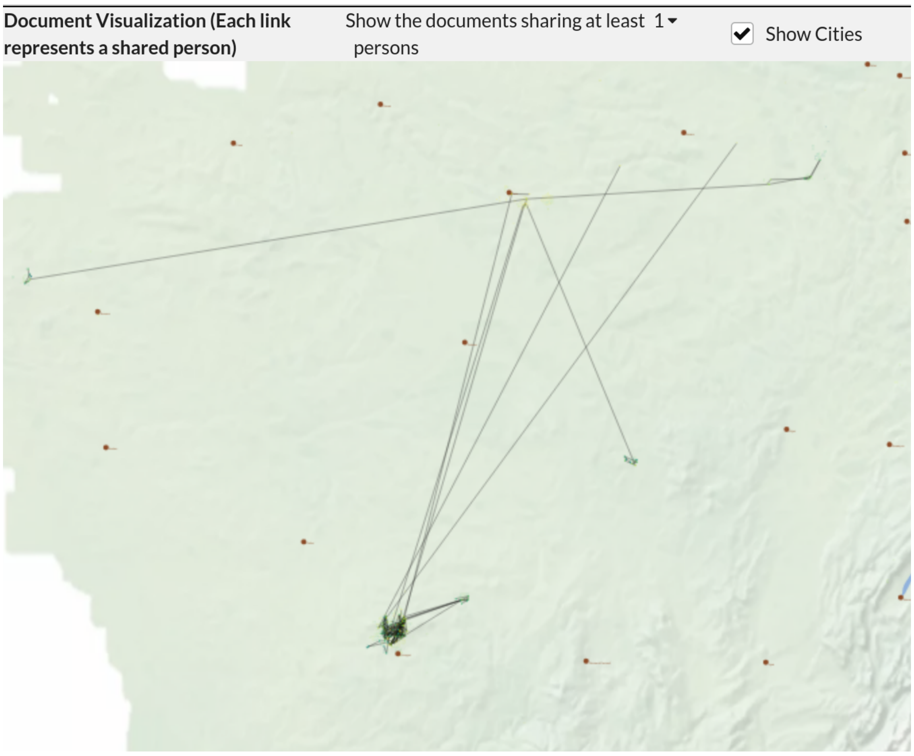
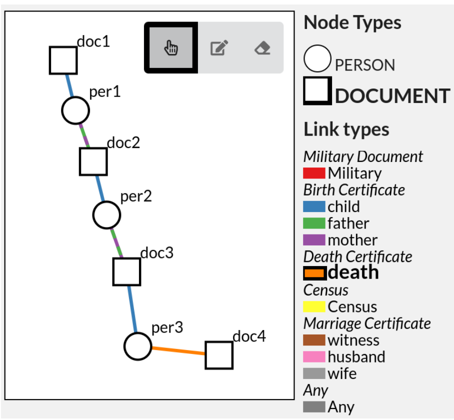
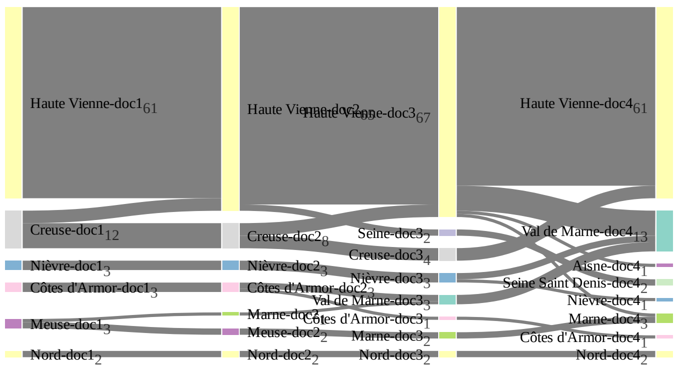


Figure 4.13 – Map of the migrations in France which occurred across several generations.



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



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

Figure 4.14 – Migrations across departments over three generations

Sankey graphs related to *departments* (Figure 4.15). 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 stayed in the same place from their birth to their death, many individuals moved of departments. It thus confirms the hypothesis of our collaborator.

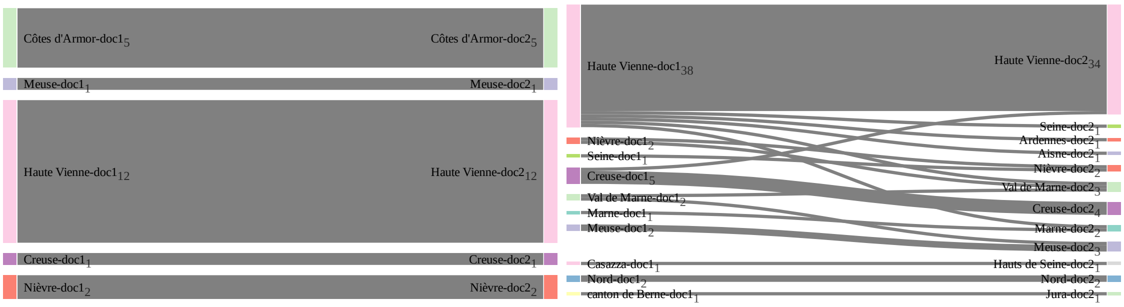


Figure 4.15 – Sankey diagrams showing the migration of people in the 18th (left) and 19th (right) centuries, extracted from their birth and death places.

4.5.3 Marriage Acts in Buenos Aires (#3)

I present in this subsection how ComBiNet has been used to detect erroneous encoding during the annotation process of the marriage acts of our collaboration #3. The 1381 documents mention 6659 individuals, who can have the same name, especially between fathers and sons in this period and region as specified by our collaborator. During the annotation process, the historian and his collaborators gave identifiers to the persons mentioned in the documents—which is typically part of the annotation procedure. However, in the case of homonyms, it can be hard to know if some mentions between different documents refer to the same or different persons. Historians cross the information contained in the different documents to disambiguate the persons [247], but errors can easily be made, i.e., giving the same identifier to different persons or giving different identifiers to the same person. I used ComBiNet in collaboration with researchers of this project to detect erroneous patterns and help them refine their encoded data. For this, we can find the persons mentioned in two acts either as *husband*, *wife*, or *witness* with a time difference of 70 years or more. Such person nodes in the network represent with almost full certainty two different persons who lived in different generations. We constructed a request to find this pattern with the visual query view and added the time constraint between the two marriage acts with the Cypher textual input. Figure 4.16 shows the visual query constructed (left), the bipartite view with the persons and documents matching the query highlighted (middle), and the table listing all the documents having mentions of people with erroneous identifiers (bottom right). The table permits to browse through all person nodes (29 have been found) that correspond in fact to more than one person and to the documents which contain the wrongly given identifiers. Using the exporting capabilities, our collaborator has exported the occurrence table indicating the pairs of documents (with their identifiers) mentioning two different persons who have been given the same identifier. Using this table, he has been able to rapidly correct those errors in his annotation framework.

4.5.4 Sociology Theses in France

I describe in this fourth use case how ComBiNet can be used to answer questions about the thesis in France between 2016 and 2022. Indeed, some sociological datasets made of documents

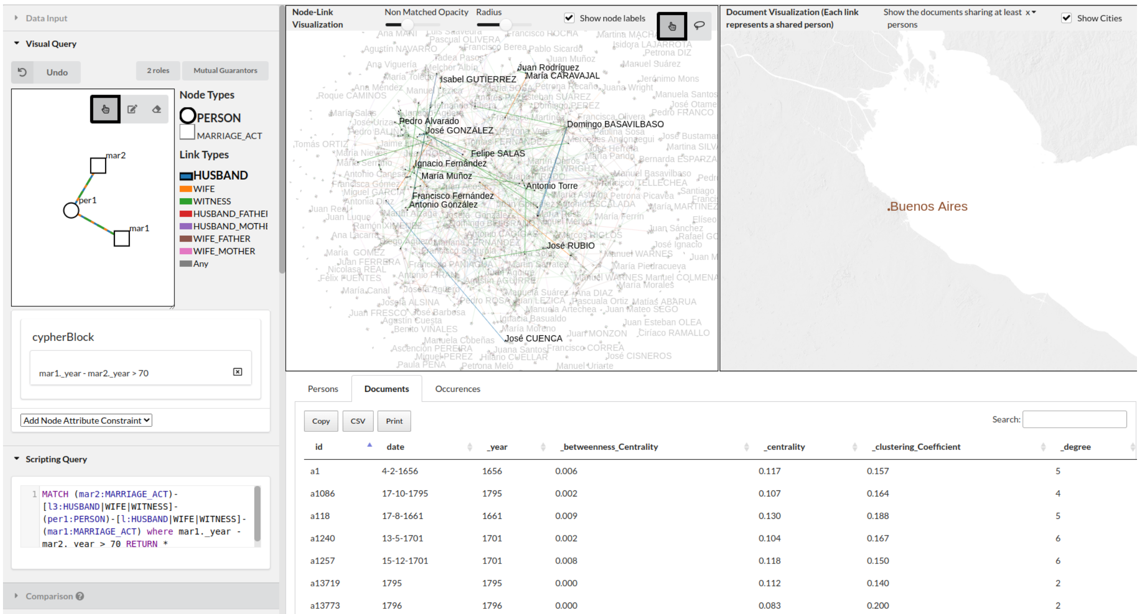


Figure 4.16 – CombiNet used to request persons appearing as husband, wife, or witness in two marriages that occurred 70 years apart or more.

can also be well modeled as bipartite multivariate dynamic networks, like 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*), reviewers (*rapporteur*), and jury president (*président de jury*). We present here an exploration of the data by ourselves using CombiNet. A first look at the graph measures tells us that 896 theses have been defended in sociology in France between 2016 and 2021 in France, with 2453 persons included in the defenses (see Figure 4.17 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 us to center the view on specific parts of the network. The map view reveals that theses have been defended all around France, even though the majority are defended in Paris. This is confirmed by a look at the distribution of the cities (Figure 4.17 bottom right): around half of the defenses are in Paris, compared to the rest of the country which is more or less homogeneous. By setting the threshold to link creation to one (meaning that a link is created between two documents if they mention at least one common person), a lot of links are created as seen in Figure 4.17 (right). It means that a lot of thesis defenses include reviewers and juries from different cities.

Let's now try to answer an interesting question: "Do reviewers and jury presidents often ask thesis directors to be reviewers and jury presidents in their turn for another thesis where they are directors ?". For this, we can construct a visual query representing this pattern by creating two person nodes and two document nodes, and by connecting them with two president links and two reviewers or jury director links in a symmetrical way, as shown in Figure 4.18 (right).

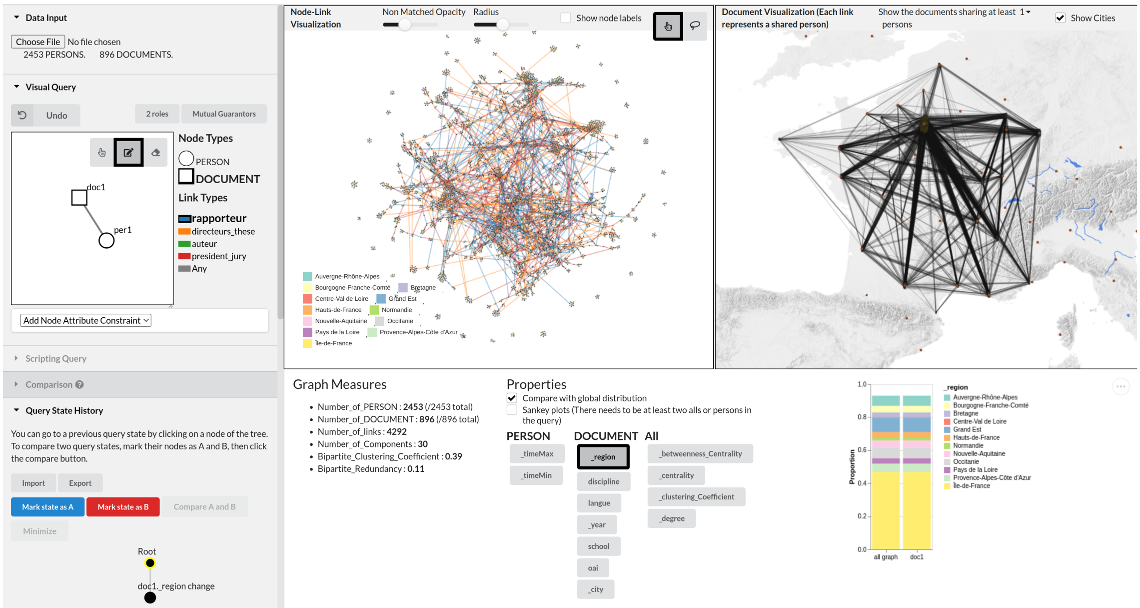


Figure 4.17 – ComBiNet used for exploring theses of sociology defended in France between 2016 and 2021. The bipartite and map views show an overview of two visions of the network. The user selects the *region* attribute, showing the geographical distribution of the defended thesis.

The occurrence table tells us that this pattern has been found 76 times in the network, meaning that this is a recurrent behavior. We are now interested in characterizing the thesis occurring in this pattern, by their regions. We can look at the *city* attribute distribution for this thesis by selecting it in the attribute view as shown in Figure 4.18 (bottom right). We can first see on the map that this pattern occurs mainly in the biggest cities of the country. By selecting the Sankey view option, we can investigate if this pattern occurs between thesis defended in different regions or if it occurs mainly in the same ones. We learn that it depends mainly on the regions: in Bourgogne-Fanche-Comté 26 out of 29 theses are connected with the thesis of another region. On the contrary, in *Occitanie* it is the case for only 4 out of 17. On average, we can see that this pattern occurs a lot for theses of the same region. In Ile-de-France, it is the case for around half of the thesis (28/50). This exploratory analysis shows that ComBiNet can be used to explore and gain insight into this type of sociological dataset.

4.6 Formative Usability Study

I performed a formative usability study with two historians and one expert in visualization. I had 3 meetings with each and gave them control of the tool to see if they could use it to explore their data—the visualization expert used the interface with the dataset of construction contracts #1—and performed queries and comparisons. At the first meeting, I explained to them the panels of the system and each feature. During each session, they were free to explore the data as they wanted. If they were stuck using one feature, I helped them by explaining

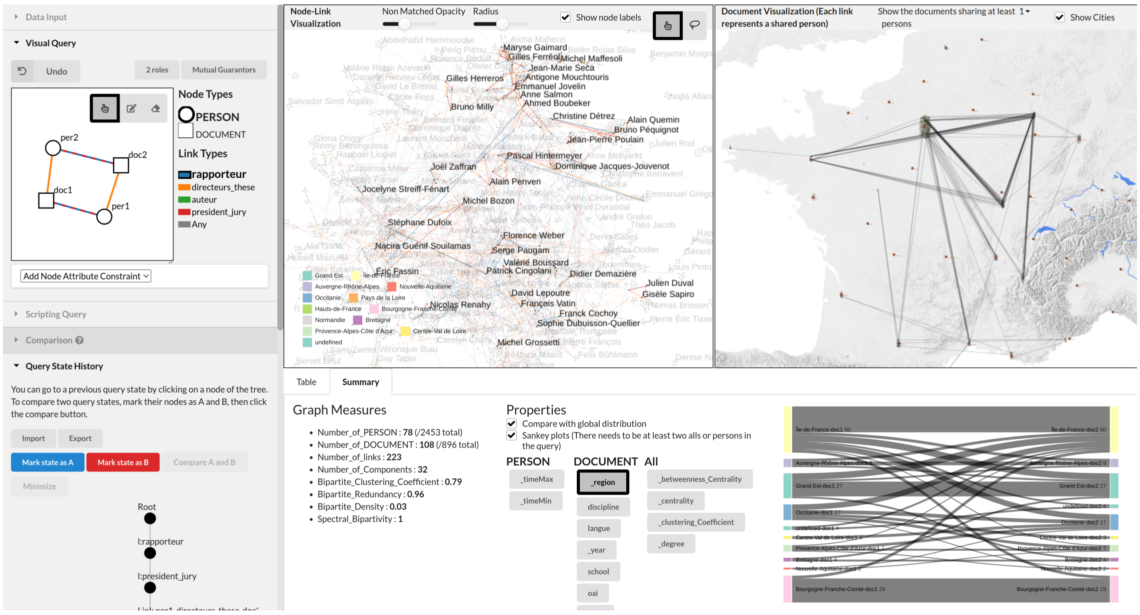


Figure 4.18 – Sociology thesis dataset explored with ComBiNet. The user constructed a visual query to see if there are symmetrical relationships between thesis directors and reviewers (or jury directors). The *region* attribute is selected with the Sankey option, letting the user see if there are correlations between the regions of the thesis found in this pattern.

how to use it. If they seemed to not know what to do next, I asked them to answer a specific question on the data (for example "can you find the persons who collaborated in more than two contracts between 1711 and 1714 in Torino"). At each meeting, I asked them to speak aloud, commenting on their aims and actions. At the end of each session, they reported their general feedback, what they did not like or understand, and what other features they would like to have to explore their data.

I improved the system and made the changes asked by the users before setting up new appointments. This usability study led to the redesign of some core features, like the activation of the comparison mode which is now started by first marking the state nodes in the provenance tree. It also led to the implementation of new features, such as the person and document tables (which are updated after each query), the persistent selection of nodes across the two views and the tables, and the undo feature for visual queries. At the final meetings, the three users were able to perform exploration, queries, and comparisons to answer socio-historical questions by themselves.

4.6.1 Feedback

All three users liked the table views and were exploiting them to study in depth who were the person and documents found in their specific queries. Both historians liked the Sankey diagram of the attributes, allowing them to see the evolution of distributions and answering several of their questions. Our collaborator of the use case #2 was making sense of it by linking the migration patterns she was seeing in the Sankey diagram with specific persons of the dataset she knew in depth. She was also curious about other migration patterns she was not aware of and wanted to know who these persons were, the system allowing her to select them and follow a deeper exploration. One other historian outside of our direct collaborators liked the overlay of node attributes on the bipartite and map views, and the distribution plots. She said: "With this data model, even if historians realize the structure of their network does not allow them to answer their research questions, they can still visualize and compare attributes of documents and persons using the visualizations, which is always useful in quantitative history".

4.7 Discussion

I discuss in this section several points of potential limitation for the system.

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 [166].

Network Modeling In chapter 3, I claimed that VA interfaces for HSNA should not only focus on answering high-level analysis questions but also support social historians in their annotation (step 3) and modeling process (step 4). I showed that ComBiNet can be used to reflect on the annotations and detect erroneous patterns (precisely in §4.5.3.), leveraging data modeled as bipartite multivariate dynamic networks. Concerning the modeling step, bipartite multivariate dynamic networks model the sources in all their complexity, and analysts may want to project the network to have a specific vision of the data to answer a precise question. ComBiNet allows focusing on specific parts of the network (one type of role, the time, the locations, or specific attributes) with the help of filters, which return subgraphs highlighting certain aspects of the data. However, it does not allow to do more complex projections—such as representing only the persons in a projected network—in part due to the limitations of Neo4j [162].

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

Generalizability. The system has been designed specifically for bipartite multivariate dynamic networks, which models well a diversity of historical sources we encountered via our collaborations: marriage acts, birth/death certificates, construction/work contracts, census, and migrations forms. Moreover, bipartite multivariate dynamic network can also be used to model other similar data types, such as scientific publications or thesis data. However, other kinds of historical textual data exist where documents can mention each other, such as in private letters for example. The model and interface would need to be slightly modified to take into account document-to-document links for these datasets. Bipartite networks are also used in various other disciplines, such as biology [130] and chemistry [133]. ComBiNet could be extended to these other application domains, in particular by modifying the map view to show other location data related to the entities of the network, or removing it altogether if it makes no sense for a particular domain.

4.8 Conclusion and Future Work

I presented in this chapter ComBiNet, a VA system for exploring social networks modeled from historical textual sources, primarily aimed at social historians. It relies on modeling documents as bipartite, multivariate, dynamic, geolocated social networks where persons are linked to documents or events using typed links that express roles. With this data model, ComBiNet lets historians explore a concrete representation of their annotated documents (i.e., the model expresses the *reality* of the documents, without the use of projections or distortions) with *traceability* to the original sources and *simplicity*. Historians can hence reflect on their encoding process (step 3 of the HSNA workflow as described in chapter 3) and answer their socio-historical questions using 1) dynamic queries on the network structure and attributes to highlight groups of interests (step 4 of the HSNA) or erroneous patterns, and 2) visual comparisons to contrast selected groups according to their structure, location, time, or any other attribute. The results can be visualized as a bipartite node-link diagram, a geographical map, graph measures, and distributions of values for the attributes. I have shown that complex explorations and analyses were easy to perform—hence giving a proof-of-concept answer for **Q2**—, and validated our approach by first describing four use cases among several other projects we are collaborating with and by performing a formative usability study showing that after many improvements the system is usable by social scientists.

By specifying a unifying data model and novel high-level visual and interactive mechanisms for comparing topology, attributes, and time, social scientists were able to correct their data more easily with exploration and querying error-induced patterns. Thanks to the document-centered model, it was easy for them to trace back the errors and inconsistencies to the sources for corrections. With the same representation, they were able to operate explorations and analyses using easy-to-use interactions implemented in ComBiNet such as coordinated views, visual querying, and comparison. Using these mechanisms, social scientists performed visual exploratory analyses of their network based on topological and attribute descriptions and comparisons of subgroups of interests—between them or the overall network. This methodology allows them to either ground or refute their hypotheses in network measures and attribute distributions, or to generate new ones from new insight revealed thanks to the exploratory and interaction mechanisms.

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

For future work, ComBiNet could be extended to support more SNA measures and computations such as clustering; it would create a new attribute containing a cluster identifier. The interface currently proposes two layouts based on the topology and the geolocations of the entities. Providing more layout options could be interesting, especially one to highlight better the time, similar to the PAOHvis technique [234]. Finally, the interface could in the future make suggestions on the query construction process based on frequent subgraphs similar to VERTIGo [59], and within a mixed-initiative perspective [155].

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

5 PK-Clustering: Integrating Prior Knowledge in Mixed-Initiative Social Network Clustering

Contents

5.1 Context **84**

5.2 Related Work **87**

5.2.1 Graph Clustering 87

5.2.2 Semi-supervised Clustering 87

5.2.3 Mixed-Initiative Systems and Interactive Clustering 88

5.2.4 Groups in Network Visualization 88

5.2.5 Ensemble Clustering 88

5.2.6 Summary 89

5.3 PK-clustering **89**

5.3.1 Overview 89

5.3.2 Specification of Prior Knowledge 91

5.3.3 Running the Clustering Algorithms 91

5.3.4 Matching Clustering Results and Prior Knowledge 92

5.3.5 Ranking the Algorithms 93

5.3.6 Reviewing the Ranked List of Algorithms 94

5.3.7 Reviewing and Consolidating Final Results 95

5.3.8 Wrapping up and Reporting Results 99

5.4 Case Studies **100**

5.4.1 Marie Boucher Social Network 100

5.4.2 Lineages at VAST 102

5.4.3 Feedback from Practitioners 104

5.5 Discussion **105**

5.5.1 Limitations 106

5.5.2 Performance 106

5.6 Conclusion **107**

As discussed in chapter 1, most SNA tools propose the computation of network measures and data mining algorithms such as clustering. Yet, social scientists are not always in a position to use them efficiently due to interpretability issues and can become frustrated with automatic results if they do not match their prior knowledge. In this chapter, I address **Q3** by proposing a mixed-initiative approach for network clustering based on the prior knowledge of social scientists, the consensus of algorithms, and exploration capabilities. In this framework, social historians are able to leverage algorithmic power in support of their analysis through interaction while limiting the introduction of bias with reports of actions leading to the final clustering. The system focus on traceability, simplicity, and document reality principles, by respectively reporting the choices leading to the constructed clusters, simple interaction mechanisms, and by leveraging bipartite multivariate dynamic networks as a data model.

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

5.1 Context

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 (see chapter 3). Before the start of the cluster analysis, a great deal of effort goes into analyzing other data and gathering knowledge—which I refer to under Prior Knowledge (PK) in the rest of the chapter. Social scientists study in great detail 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 hypotheses and conclusions. When the network is small, with 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 [225], Gephi [18], Python with NetworkX [101], or Pajek [171]. To compute clusters, they follow a strained process: they select and run algorithms provided in the tool and then try to make sense of the results (see Figure 5.1 left). When they are not satisfied or unsure, they iteratively tweak the parameters of the algorithms at hand, run them again, and hope to get results more aligned with their prior knowledge. This analysis process is unsatisfactory for three main reasons:

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

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 results [131]. In SNA, this means, for example, that the same social network where links represent a global notion of proximity could be clustered to find families, friend groups, or business relationships. One partition is not necessarily better than another one but depends on the purpose of the analysis. This issue increases the need for interactive tools, which let the user specify which type of partition is expected.

To address those issues we propose a novel approach, called PK-Clustering, which allows social scientists to iteratively construct and validate clusters using both their *prior knowledge* and consensus among clustering algorithms. A prototype system illustrates such an approach, and provides a concrete example of a solution to **Q3** in the context of social network clustering: how to design VA tools and interactions that leverage algorithmic power but keep historians in control of their analyses and biases?

The proposed approach includes three main steps (see Figure 5.1 right):

1. *Specify Prior Knowledge (PK)*. Users introduce their prior knowledge of the domain by defining partial clusters. The tool then runs all available clustering algorithms.
2. *Consolidate expanded PK clusters*. Users review the list of algorithms, ranked according to how well they match the prior knowledge. They compare results and consensus, then accept or ignore suggestions to expand the prior knowledge clusters
3. *Consolidate extra clusters*. The tool suggests extra clusters on unassigned nodes. The user reviews the consensus on each proposed cluster and 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 [215] and inspired by the idea of letting the user collaborate with the machine to reach specific goals [116], the proposed approach follows a user-initiated mixed-initiative [116] visual analytics process.

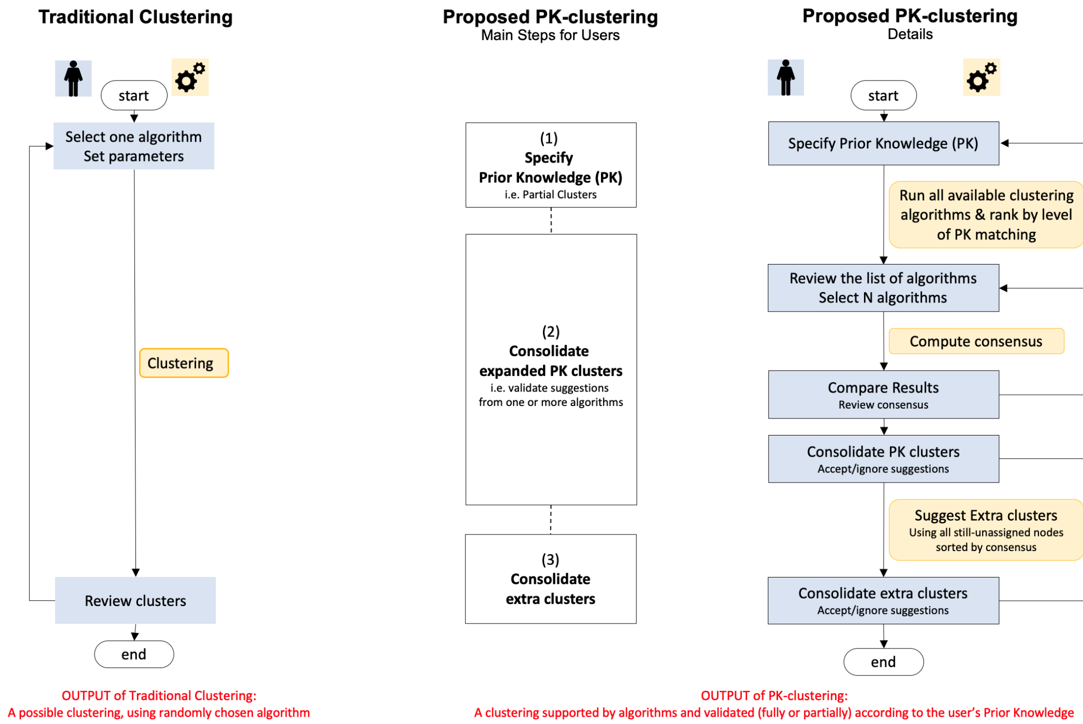


Figure 5.1 – Process of traditional clustering (left) and our PK-Clustering approach (middle and right). The output of traditional clustering is a possible clustering, using an algorithm among many choices. The output of PK-Clustering is a clustering supported by algorithms' consensus and validated (fully or partially) according to the user's PK.

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

Contrary to a current trend [165], 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 fits 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. PK-Clustering constitutes a new approach for social network clustering, that I demonstrate with a concrete prototype and validate with two case studies.

5.2 Related Work

PK-Clustering relies on several families of clustering methods and the visualization and exploration of their results. I first describe a brief overview of clustering for graphs, as well as semi-supervised methods, then several works in the literature related to VA: interactive clustering, groups in network visualization, and ensemble cluster visualization.

5.2.1 Graph Clustering

One of the main properties of social networks is their community structure [91] which reveals group relationships between nodes, known as communities or clusters, which have a higher density of edges compared to 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 an abundant and growing literature on clustering methods to find these communities for social networks. The majority of the research is made only on topological algorithms, i.e., algorithms which use only the structure of the network to find clusters. [83] 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 a 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-world phenomena are often best modeled with bipartite graphs, also known as 2-mode networks. It is the case for social historians, who often build their networks from raw documents containing mentions of people, as discussed in chapter 3. Several algorithms exist for bipartite graph community detection [7].

Moreover, recent new approaches try to use the attributes of the nodes [251] and the dynamic aspect of the networks [197] to find more relevant communities. Some toolkits offer a large number of algorithms; for example, the Community Discovery Library (CDLIB) [198] implements more than 30 clustering methods with variations inspired by 67 references.

5.2.2 Semi-supervised Clustering

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

Semi-supervised clustering can be divided into constraint-based and seed-based clustering. The former includes must-link (ML) and cannot-link (CL) constraints [19, 238]. 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 [19], Fuzzy-CMeans [24], hierarchical clustering [28], Density-Based Clustering [144], and graph-based clustering [238]. Shang et al. [212] use a seeding then expanding scheme to discover communities in a network. Their clustering method considers links as documents and nodes as terms.

Swant and Prabukumar [205] review graph-based semi-supervised learning methods in the domain of hyperspectral images. Vertices of the graph represent items that may be labeled, while the edges are used to specify the similarity among the items. The technique classifies unlabelled items according to the weighted distance from the labeled items.

5.2.3 Mixed-Initiative Systems and Interactive Clustering

Introduced by Horviz [116], mixed-initiative systems are "interfaces that enable users and intelligent agents to collaborate efficiently". Several Visual Analytics systems are based on mixed-initiative interactions, e.g. [51, 155, 239, 253], in particular interactive clustering systems.

PK-Clustering is an interactive clustering system. A review by Bae et al [15] shares our concerns: "Real-world data may contain different plausible groupings, and a fully unsupervised clustering has no way to establish a grouping that suits the user's needs because this requires external domain knowledge." Interactive clustering systems aim at producing visual tools that let users interact and compare several clustering results with their parameter spaces, making it easier to find a satisfactory algorithm for a particular application. Several such systems exist (e.g. [45, 153]) but few deal with network data. These systems adapt one algorithm to become interactive using some type of constraints. Instead, our approach applies ML/CL constraints on a wide variety of existing algorithms, providing richer algorithms and control than the reviewed systems.

5.2.4 Groups in Network Visualization

To assess the quality of clusters in networks, 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. [237]. 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 [121]. 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 [203], using new tasks assessing group-level understanding.

Holten [114] proposes edge bundling on compound graphs. He bundles together adjacent edges, making explicit group relationships at the cost of losing the detailed relationships. A good example of manual grouping and tagging is SandBox, which allows users to organize bits of information and their provenance in order to conduct an analysis of competing hypotheses [249]. A lot of work has also been done on the visualization of categorical variables in tabular data [97, 135], which is similar to the notion of groups in networks.

5.2.5 Ensemble Clustering

In the context of machine learning, an ensemble can be defined as "a system that is constructed with a set of individual models working in parallel whose outputs are combined with a decision fusion strategy to produce a single answer for a given problem" [240]. Several strategies exist for combining multiple partitions of items in a clustering setting [223]. Concerning visualization research, Kumpf et al. [137] consider ensemble visualization as a sub-field of uncertainty visualization, for which some surveys exist [29,154]. 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 representations of clustering memberships for each member. Our system provides a consensus-based interactive strategy that takes into account users' prior knowledge instead of relying on mathematically defined optimal assignments only.

5.2.6 Summary

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

Standard clustering. One algorithm is picked with a set of parameters and the user checks 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 users in the analysis loop and allow them to have a better impact on the final community detection result.

5.3 PK-clustering

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

PK-Clustering is not tied to any specific network representation technique and could be used to augment any of them. Our prototype is implemented in the PAOHVis tool [234] 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, I describe each step in more detail, illustrated with screen samples taken during the analysis of a small fictitious network.

5.3.1 Overview

In PK-Clustering the user and the system take turns to construct and validate clusters. The process involves three main steps, each with several activities (see Figure 5.1 right). The blue boxes describe the user activities while the yellow boxes describe the system activities. After loading the dataset, the process is as follows:

(1) Specify Prior Knowledge (PK).

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

(2) Consolidate expanded PK clusters.

3. Users review the ranked list of algorithms. They can see if the algorithm results match the PK completely, partially, or not at all. Information about the number of clusters generated by each algorithm is also provided. Users select the set of N algorithms they think are the most appropriate.
4. The consensus between the selected algorithms is computed and visualized next to the graph visualization (in the PAOHVis display in our prototype)
5. Users review and compare the suggestions made by the algorithms to expand the PK-groups, i.e., the groups defined by the PK, 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 network visualization, and the results of the clustering algorithms to make their choices.

(3) Consolidate extra clusters.

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

At 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 at all, and accept all consensual suggestions without reviewing them in detail. 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, as it often happens that social scientists do not have a strong opinion on the group appartenance of some individuals. 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 the review of 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, and 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 detail.

5.3.2 Specification of Prior Knowledge

Users start the process by expressing their PK as a set of groups. Each group contains the node(s) that the expert is confident belong to the defined group. In the case of Figure 5.2, each of the two prior knowledge groups contains two nodes, and it specifies that the user is expecting to see at least two clusters, with the first two people in a blue cluster A, and the other two in a red cluster B. This representation expresses *must-link* and *cannot-link* constraints described in §5.2.2 in a simple visual and compact form. It is not required to specify all binary constraints because the information is derived from the prior knowledge groups.

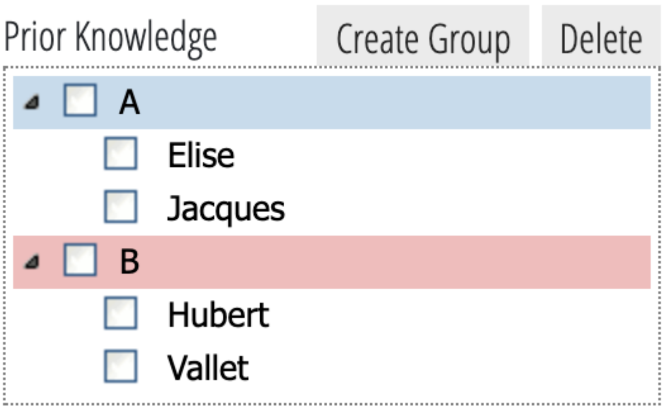


Figure 5.2 – Prior Knowledge specification, the user-defined two groups composed of two members.

5.3.3 Running the Clustering Algorithms

The prototype includes 11 algorithms taken from three families:

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

Topology-based algorithms. Most of the clustering algorithms consider only the graph topology and optimize a topological measure such as *modularity* [37]. 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 nodes of the same group compared to the rest of the graph.

Propagation/Learning-based algorithms. Semi-supervised machine learning algorithms learn from incomplete labeling 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' PK groups in its clusters computation. By design, this type of algorithm will always provide a perfect match with the PK, even if the PK does not make much 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 [63] and Bipartite Modularity Optimisation []. Since the majority of community detection algorithms are for unipartite graphs, the system performs a projection into a one-mode network [254]. Basically, each pair of nodes that are in the same hyperedge are connected together in the resulting graph, with a weight being the number of shared hyperedges [].

Some algorithms require parameters to be specified. We do not force the user to specify values for all the parameters, when possible, we infer them from the PK-groups. For instance, instead of using an arbitrary default for the number of expected clusters k in k-means clustering, we run the algorithm several times with a value of k from the number of specified PK-groups to this number plus two. The strategy of using several parameter combinations for the same algorithm is often used in ensemble clustering to increase the number of different clusterings. However, the number of parameter combinations can be extremely high. The research field of *visual parameter space exploration* (see e.g., [208]) 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, the resulting clusterings are matched with the PK and ranked by how interesting their results are likely to be for the user.

5.3.4 Matching Clustering Results and Prior Knowledge

Once a clustering is computed, we want to know how well it is compatible to the PK, and if possible, match every PK-group with a specific cluster. We use the *edit distance* to measure this matching, as its computation allows us to directly link each PK-group to a specific cluster. Given two partitions, the edit distance is the number of single transitions to transform the first partition into the second one. For example, the edit distance between the two partitions of 4 nodes $P_1 = \{\{1,2,3\},\{4\}\}$ and $P_2 = \{\{1,2\},\{3,4\}\}$ is 1 because moving the node 3 from the first to the second set of P_1 would transform it into P_2. A clustering can be seen as a partition since every node has a label, but the PK can only be seen as a partial partition because only some nodes are labeled. We say that the edit distance between the PK and a clustering is 0 if every group of the PK is a subset of an exclusive cluster, i.e., if every person of a PK-group is retrieved in the same cluster, with no overlaps. Thus, we define the edit distance as the number of node transitions between the groups of the PK to get to the state where each group is a subset of an exclusive cluster. More formally, we can express this as a maximum weight bipartite matching problem [132], where the PK $PK = P_1, \dots, P_n$ and a given clustering $C = C_1, \dots, C_n$ constitute the bipartition (PK, C) of a bipartite graph $G = (V, E)$. A link is created if a PK-group and a cluster share nodes, with a weight equal to the number of shared nodes, giving the following weight function:

$$w(PK_i, C_i) = card(PK_i \cap C_i) \tag{5.1}$$

We then need to find a matching M of maximum weight w, with

$$w(M) = \sum_{e \in M} w(e) \tag{5.2}$$

This can be done with the Hungarian method [136]. The Matching gives the correspondence between the PK-groups and the clusters computed by a given algorithm, and the edit distance ED is given by the number of nodes specified in the PK minus the total weight of the matching:

$$ED = \sum_{i}^{n} card(PK_i) - w(M) \tag{5.3}$$

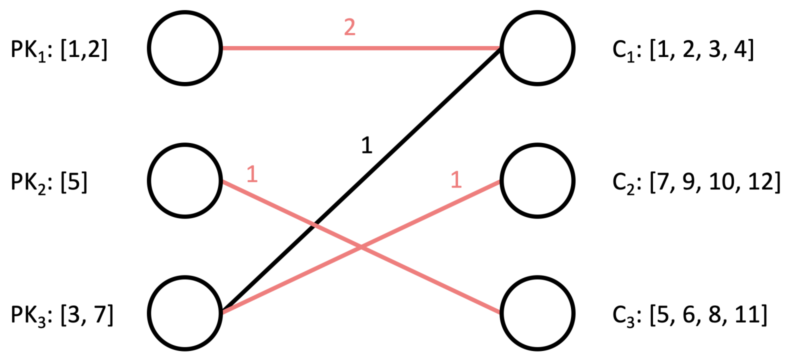


Figure 5.3 – Red edges represent the prior knowledge matching

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 Figure 5.3. 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 a unique cluster, with no overlap.

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

5.3.5 Ranking the Algorithms

The algorithms are ranked by their degree of matching with the PK, 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 PK 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 users to reconsider their PK, as it does not match the data encoded in the network.

5.3.6 Reviewing the Ranked List of Algorithms

Once the clustering algorithms have been matched with the PK, users can review the list of algorithms, ranked by how well their results match the PK. Figure 5.4 shows two modalities to visualize the ranked list (individual nodes and aggregate representation). I will describe in details the first modality, which shows individual nodes as small colored circles:

Each row is an algorithm, and the algorithms are grouped by family. On the right of the name of the algorithm, there is a representation of the clusters that best match each of the PK-groups. Figure 5.4 shows first 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 an X for no match. Additional nodes in the cluster are represented as white dots with a number next to it. On the right we see how many other clusters (if any) have been found by the algorithm—also represented as white dots with a number next to them.

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. The top four algorithms match the PK perfectly, while the following one *fluid_k4* has a partial match. At the bottom, an algorithm (*label-propagation*) has no match.

The alternate modality of representing the matches (shown at the bottom of Figure 5.4) 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 larger networks because it allows users to see the results in a more compact way.

Once users have reviewed the list of algorithms they can review the 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 criteria: the *degree of matching* (i.e., they can choose to look at algorithms with no match to challenge their prior knowledge); the *algorithm type* (the user may prefer an attribute-based algorithm, rather than one based on topology); the *size* of the matched clusters; or the number and size of *other clusters* found by the algorithm.

PK-Clustering expresses its prior knowledge through *must-link* and *cannot-link* constraints. However, at this stage, the user can decide to use this expressive power as strong constraints—

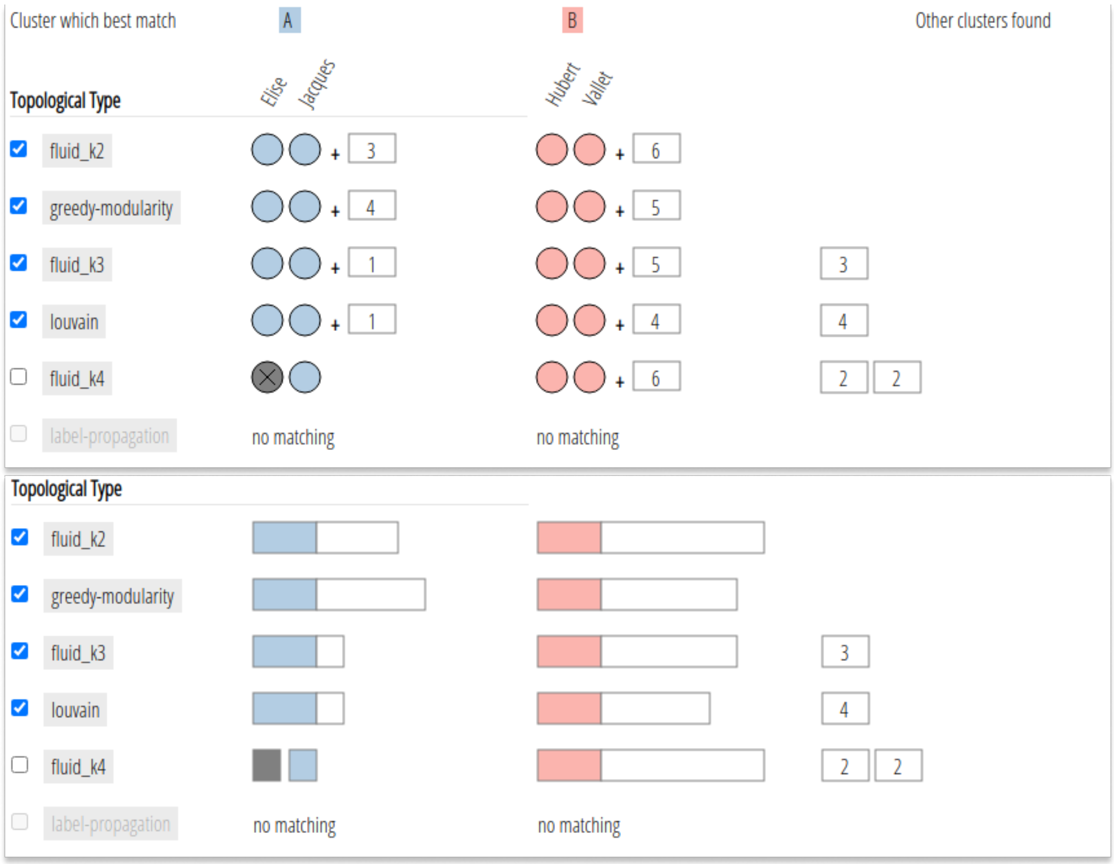


Figure 5.4 – Two different modalities for the ranked list of algorithms. Top: persons are shown as circles. Bottom: aggregated view. Colors indicate the matching group. Gray indicates no match. White indicates extra nodes or clusters.

only selecting algorithms that match all of them—or as weak constraints—to explore clustering results that support most or some of them. Our historian colleagues have used both, either to cluster a well-understood dataset with strong constraints or to generate hypotheses on less known ones.

5.3.7 Reviewing and Consolidating Final Results

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

Reviewing Results of a Single Algorithm

By clicking on an algorithm name the results of that algorithm are displayed in the PAOHVis view (see Figure 5.5). In this view, each line corresponds to a person in the graph, and each vertical line represents a hyperedge connecting them [234], in a way visually similar to the UpSet representation [150] but semantically different. Alternative graph representations are available as well—such as node-link diagrams—but the PAOHVis view is well adapted to PK-Clustering.

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

Comparing Multiple Algorithm Results

From the ranked list of algorithms, users can select a set of algorithms and click the large green button to review and compare the selected algorithms in the PAOHVis view (see Figure 5.5 By default, the PAOHVis view groups the names using the clusters of the 1st algorithm, but on the left of the node names now appears complementary information about the results of all the selected algorithms.

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

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

In Figure 5.5 PK_louvain is selected as the base algorithm for the grouping of names in the list. We see that there is a 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 on 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.

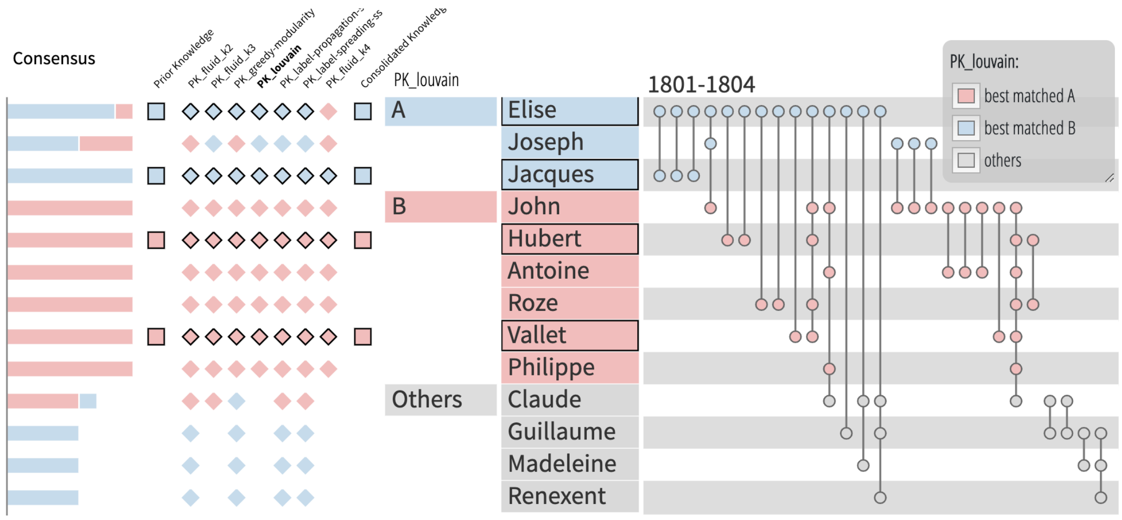


Figure 5.5 – Reviewing and comparing results of multiple algorithms. One algorithm is selected to order the names and group them, but icons show how other algorithms cluster the nodes differently, summarized in the consensus bar on the left.

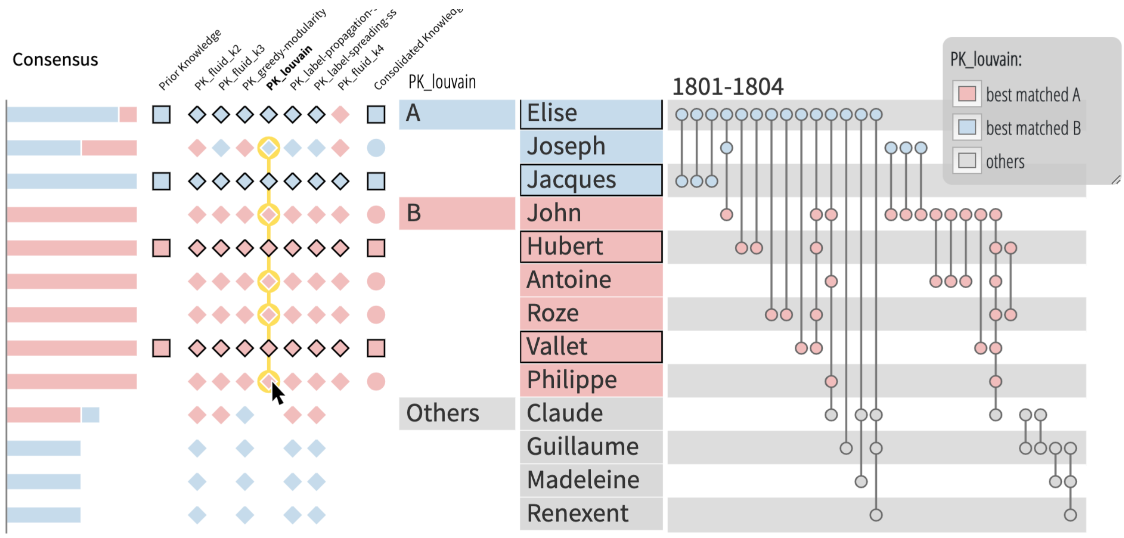


Figure 5.6 – The user quickly drags on consecutive icons (in yellow) representing the suggestions made by one algorithm to validate node clustering. Once the cursor is released the validated nodes appear as squares icons in the Consolidated Knowledge column.

Consolidating the prior knowledge clusters

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

Partial Copy. By clicking on one of the icons or dragging the cursor down on a set of icons, users validate the suggestion(s) of an algorithm, adding colored squares in the consolidation column. Once this validation is done, the squares do not change color anymore and represent the user's final decision (unless changed manually again). Figure 5.6 shows how a user drag-selects a set of diamonds in the column PK_fluid_k4. They are connected by a yellow line, which appears while dragging over the icons. When done the status of the nodes in the Consolidated Knowledge column (rightmost) will change to square.

Consensus slider. Users can set the consensus slider to a certain value (for example 4) to automatically select all nodes that have been classified in the same cluster by at least 4 algorithms. While the slider is being manipulated circles appear in the consolidated column. Then users can validate the suggestions by clicking or dragging on the circles, or by using the *consolidate suggestions* button which will validate all suggestions at once.

In summary, diamonds represent suggestions from one algorithm, circles temporary choices, and squares represent 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 a 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 users can change along the process. First, the user selects one algorithm in the PAOHVis view, and the nodes are grouped by the clusters found by the algorithm. The PK-clusters are displayed at the top, followed by *Others*, which contains everyone else. When users click on *Others*, the other clusters are displayed ordered by consensus. Since the number of clusters can be high, all new clusters appear in gray to avoid the rainbow effect. A secondary matching process matches the clusters of the current algorithm with those of all the other algorithms, one by one (similar to the matching process described in §5.3.4). Once the matching is done, the consensus of one cluster is computed as the sum of the cardinalities of the intersections between the cluster and all the other clusters of the other algorithms matched with it, divided by the number of nodes of the cluster.

When users hover over one cluster name, a new color is given to that cluster (e.g., green) and new (green) diamonds appear for each algorithm that matches the cluster and for each node that is assigned to the cluster (Figure 5.7). Users can therefore see if the selected cluster is consensual, and with which algorithms. The top part of Figure 5.7 shows the mouse pointer before hovering on cluster 2. The bottom part shows that hovering the mouse pointer over cluster 2, it changes to green and several green diamonds appear along three columns.

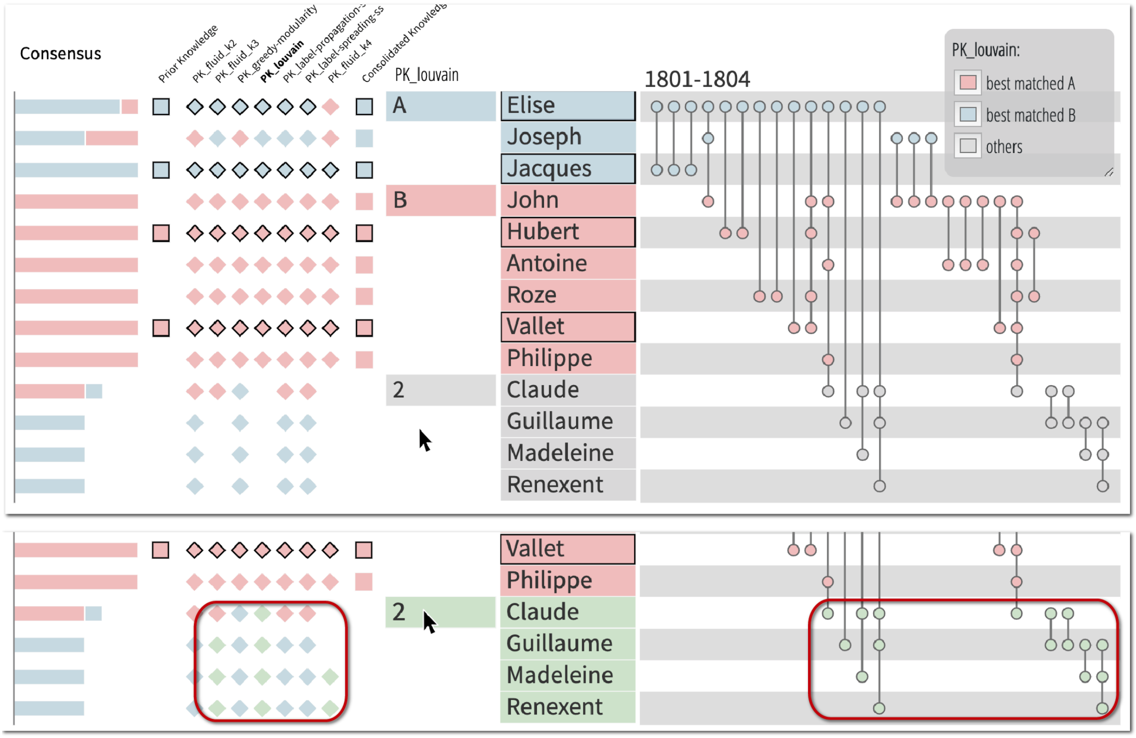


Figure 5.7 – The suggestion of extra clusters. The two PK-groups (red and blue) are validated (nodes in the consensus column are all squared). One extra cluster is proposed by the Louvain algorithm, labeled as 2. Hovering over the cluster 2, the consensus is displayed by the green diamonds. This feedback is also visible in the graph.

The evaluation of the best cluster for a node can be done using multiple encodings. The suggested clusters appear in 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 in the cluster having that color. If users are satisfied with the association proposed by the current algorithm, they can validate it by clicking on the cluster name. This will create a new group, so the user can classify the nodes into this new group, as seen before (§5.3.7): using the consensus slider, copying an algorithm result, or through manual labeling. This process is repeated for the other clusters until there are no unlabeled nodes or the user is satisfied with the partial clustering. An example of a fully consolidated dataset is shown in Figure 5.8.

5.3.8 Wrapping up and Reporting Results

At any stage of the process, the user can finish instantaneously, either by not labeling undecided nodes, selecting and validating the results of a single algorithm—as traditional approaches

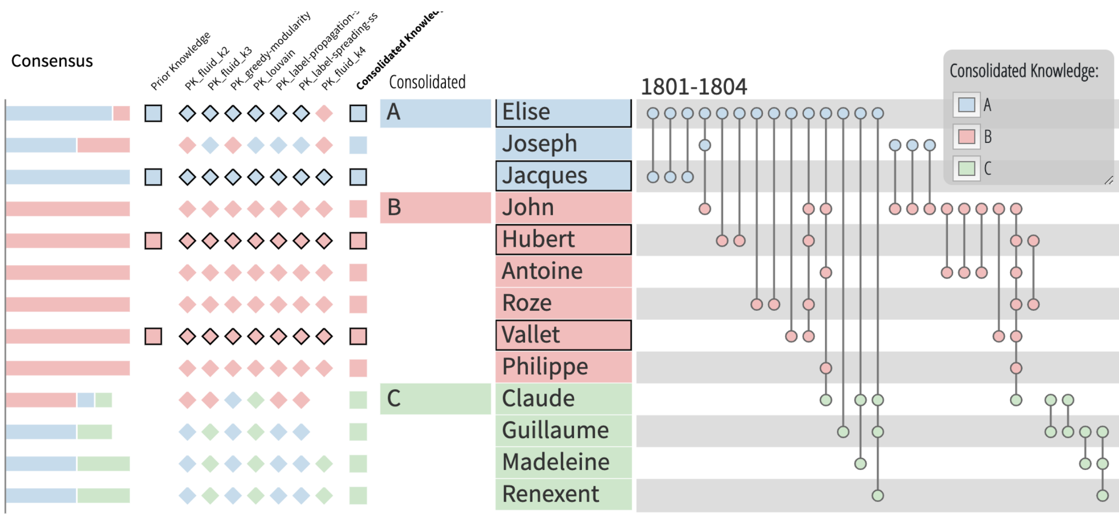


Figure 5.8 – The dataset has been fully consolidated. The persons are grouped and colored by the consolidated knowledge. The user decided to assign Claude, Guillaume, Madeleine, and Renexent to cluster C, by taking into account the graph and the consensus of the algorithms.

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 node, 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 interaction methods and can be used in a publication.

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

5.4 Case Studies

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

5.4.1 Marie Boucher Social Network

I asked one of our historian colleagues for her prior knowledge on her network about the trades of Marie Boucher [], composed of two main families: Antheaume and Boucher. Family

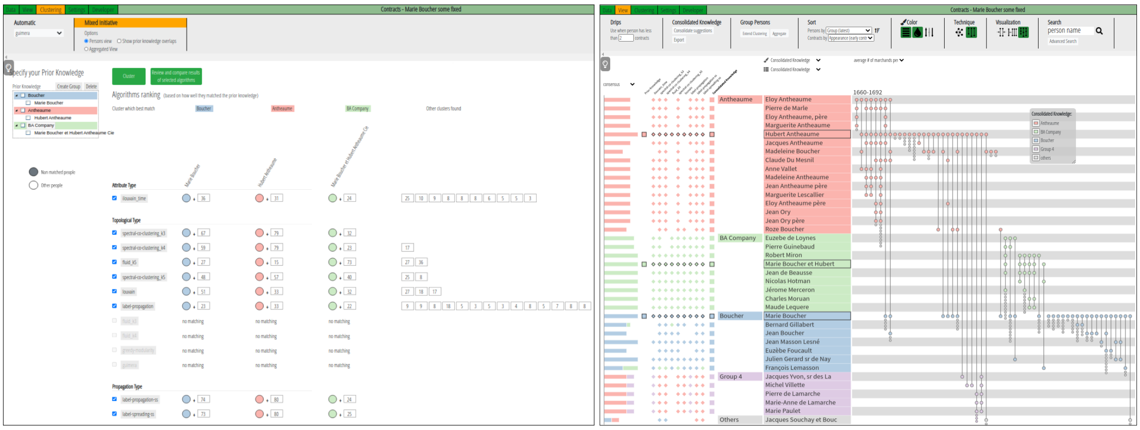


Figure 5.9 – Two main phases of PK-clustering. On the left, the user has specified the Prior Knowledge (PK) groups (top left) and then reviews the list of algorithms ranked according to how well they match the PK. On the right, the user compared the detailed results of selected algorithms and consolidated the results. From the initial specification of three groups and three people, 4 relevant clusters were obtained with 37 people in total, plus one unclassified node (*Others* group).

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 into 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 PK and started the analysis. Figure 5.9 (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 Figure 5.9 - left.) with the first algorithm listed as 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 network visualization and her own knowledge of the persons. At the end of this step, the Boucher, Antheaume, and Boucher & Antheaume groups were consolidated, but there were still several persons not labeled on the consolidated knowledge. She decided to review in more detail the clustering results using the *ilouvain_time* algorithm because of its reliance on the time attribute, and also because its results seemed good in the matching view. After clicking on the virtual group *Others*, the four other clusters computed by *ilouvain_time* appeared and were reviewed by hovering the mouse on the names of these new groups. She selected only one cluster she was confident about and consolidated it.

The final validated partition of the dataset is represented in Figure 5.9 (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 (illustrated in Figure 5.10) will help document where the final grouping came from, increasing trust with regard to her claims.

2020-07-23
Prior Knowledge: 3 groups, 3 persons
Final: 4 groups, 37 persons (out of 38 visible) consolidated as follows:
12 (32.4%) using the consensus:
  10 (27.0%) with an agreement of 8 algorithms
  2 (5.4%) with an agreement of 5 algorithms
18 (48.6%) using the result of 1 algorithm:
  18 (48.6%) with PK_ilouvain_time
4 (10.8%) manually consolidated
0 (0.0%) by clicking on the group distribution
13 algorithms considered
9 algorithms reviewed and compared

Figure 5.10 – Summary report of the consolidated knowledge for the Marie Boucher case study.

5.4.2 Lineages at VAST

In the second case study, I take the role of Alice, a VAST Steering Committee (SC) member, who participates in an SC meeting to validate the Program Committee proposed by the VAST paper chairs for the next conference. This case study illustrates the utility of PK-Clustering for finding relevant clusters in a social network modeled through documents (here scientific publications), similarly to historical sources. 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 [117] 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

Figure 5.11 – Computing the Lineages of VAST authors: Prior Knowledge from Alice and results of the clusterings matching it.

"Visualization Publications Data" (VisPubData [119]) 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 consolidated (steps 2 and 3).

Alice starts the PK-Clustering process by entering her prior knowledge, which is partial and based on two strategies: her knowledge of some areas of VAST, and the name of well-known researchers who have developed their own lineage. She runs the algorithms (Figure 5.11) and 5 algorithms produce a perfect match, acknowledging her knowledge of some areas of VAST. She then shows the results to other members of the SC who will help her consolidate the lineage clusters.

Her initial PK clusters are quickly consolidated, using Internet search to validate some less known authors. She then decides to create as many additional clusters and lineage groups as she can. For some authors, she decides to override the consensus of the algorithms. For example, she decides, and her colleagues agree, that Gennady and Natalia Andrienko should be in their own lineage group and not in D. Keim's (Figure 5.12). The history of VAST in Europe, very much centered around D. Keim and the VisMaster project [128], has strongly influenced the network structure and some external knowledge is required to untangle it.

Using the *PK_louvain* algorithm as a 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 sorts the list by group. Alice can now report her work to the whole steering committee, 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 are not always aligned with the lineages. Figure 5.12 shows a subset of the final results.

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, PK-Clustering can provide a partial but concrete answer to the problem of defining what the scientific lineages are.

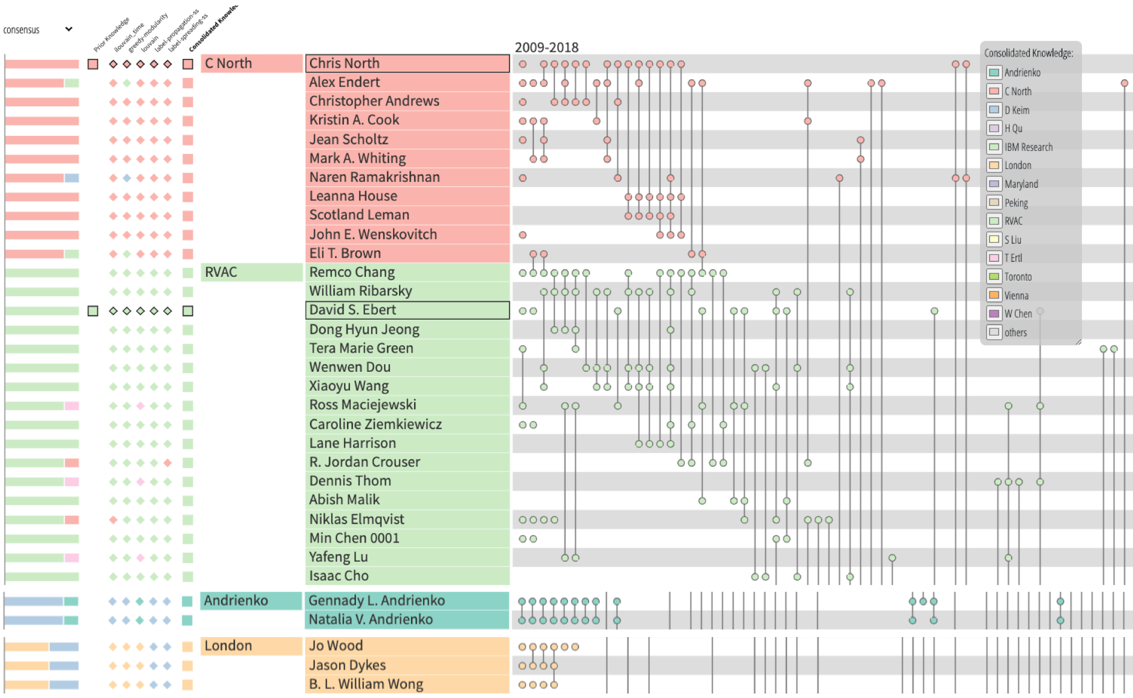


Figure 5.12 – Four consolidated groups in the VAST dataset: C North, RVAC, Andrienko, and London

5.4.3 Feedback from Practitioners

We showed the system to three practitioners and asked for their feedback through video-conferencing systems, sharing video demonstrations, and sharing our screen.

They all acknowledged the pitfalls of existing systems providing clustering algorithms as black boxes with strange names and mysterious parameters. They also agreed that the current process for clustering a social network was cumbersome when they wanted to validate the groups and compare the results of different algorithms. None of the popular and usable systems provide easy ways to compare the results of the clusterings. Usually, the analyst needs to try a few algorithms, remembering the groups that seemed good in some of the algorithms, sometimes printing the clustered networks to keep track of the different options. Still, they all confirmed that they usually stop after trying 2 to 3 algorithms because of a 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 into 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.

One other historian from the collaboration #1 described in §3.3.1 used the system in a particular way. After entering his PK on core members of families of interest and selecting algorithms matching the PK, he inspected the clustering consensus without creating a consolidated partition. He was particularly interested in non-consensual individuals that were highlighted by the colored consensus bar—particularly persons with similar numbers of algorithms classifying them into two distinct groups. He induced that those persons acted as bridges linking several different communities, here families. Instead of creating consolidated communities, he created a new group where he manually tagged all those persons acting as bridges, which was a behavior of interest to him.

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

5.5 Discussion

As presented in §5.2.6, the existing approaches to creating clusters in social networks consider three options: standard clustering, ensemble clustering, and semi-supervised clustering. The 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, social scientists are rarely satisfied with current clustering methods, in part because they run on network 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 if they want but the system 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 publications.

Bias issues are complex and the absence of ground truth limits researchers' ability to measure those biases. No approach solves all issues yet, but I believe that PK-Clustering offers a fresh perspective on those issues and will lead to results that are more useful to social scientists.

5.5.1 Limitations

Many more clustering algorithms exist and could be added. Moreover, expanding the exploration of parameter spaces for clustering algorithms seems needed. Another limitation of the current prototype is that some algorithms do not work well with disconnected components of the graph. Unfortunately, social scientists' datasets typically have many disconnected components. This issue can be mitigated by separating components into a set of connected components, run the algorithms on them, and merge the results. The 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 could also help us improve the utility of the tool as well as the provenance table and summary.

5.5.2 Performance

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

In order to improve computational scalability, progressive techniques can help to deal with larger sizes [78]. The current user interface design for PK-Clustering[1] would allow the ranked list of algorithms to be progressively updated, and users to review a few individual algorithms first while other algorithms are still running. Of course, visual scalability is also an issue with larger datasets, as the list of people also grows. PAOHVis allows groups (like clusters) to be aggregated or expanded, so we expect that users would expand clusters one by one to review and consolidate them, while also being able to review the connections between the proposed clusters. Users can also use the automated features of PK-Clustering to consolidate the nodes (e.g., selecting one algorithm based on the ranking, or using the consensus slider to consolidate all the nodes at once). Pixel-oriented visualizations [126] would facilitate the review of consensus for a large number of nodes and clusters. Classic techniques like zooming or fisheye views [120, 192] would help as long as names remain readable, which is critical to our users.

5.6 Conclusion

In this chapter, I 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, and 3) users review and compare results to consolidate the final clusters.

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

PK-Clustering follows *traceability*, *simplicity*, and *document reality* properties discussed in chapter 1 and chapter 3, by respectively providing a summary report of the actions leading to the final clustering, simple interactions, and the usage of bipartite multivariate dynamic networks as a data model. This approach is a concrete proof-of-concept solution to **Q3** in the context of clustering, as it provides a framework for social scientists, specifically historians, to follow a

[1] The web application and source code are available at `https://www.aviz.fr/Research/Pkclustering`

clustering analysis supported by algorithmic power but always in control of the decision process, through easy-to-use interactions. Clustering and social network analysis remain challenging tasks, typically without ground truth to formally evaluate the results. If PK-Clustering limits bias inherent to traditional clustering (priming bias and lack of control), the high influence of users on the decision-making may introduce other types of bias. Still, I 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.

6 Conclusion

Contents

6.1 Summary 109
6.2 Discussion 110
6.3 Perspectives 113
6.4 Conclusion 116

6.1 Summary

In this thesis, I addressed the high-level question of how VA can help social historians conduct network analysis, in their entire process, from the collection of documents to the final analysis and visualization of constructed networks. Indeed, social historians currently use visual and analysis tools to generate insights from curated networks, but the process they use is tedious and error-prone, which can result in simplification, distortions, errors, and inconsistencies [6, 146]. Moreover, current tools typically lead to qualitative descriptions of the network data [196] instead of a deep understanding of the global and local network structure supported by computation, mostly due to usability and interpretability issues. VA could therefore support historians in 1) their data preparation process, and 2) the final analysis of curated networks. Figure 6.1 shows a schematic representation of the potential place of VA in the HSNA process. From continuous discussions with social historians, I identified three principles VA interfaces

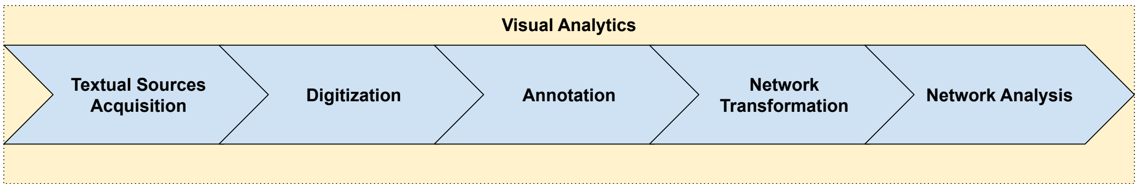


Figure 6.1 – Potential place of VA in the HSNA workflow defined in chapter 3.

should follow: *traceability*, *simplicity*, and *document reality*, to respectively ease the back and forth between the different process steps while assuring reproducibility of analyses, have expressive representations and tools that are simple enough to manipulate for social scientists, and ground results in the concrete reality of the documents, hence limiting the introduction of bias and distortion. More precisely, I tried to answer three questions with respect to those properties:

Q1: How to model historical documents into analyzable networks with the right balance between expressiveness and simplicity.

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

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

In chapter 3, I formalized the HSNA process from collaborations with social historians, into five steps: textual sources acquisition, digitization, annotation, network creation, and network visualization/analysis (see Figure 6.1); I identified recurring pitfalls for each step, such as wrongly chosen network models or named entity recognition errors. The use of VA on the whole process could limit the introduction of such pitfalls by allowing historians to reflect on the annotation, encoding, and modeling processes through visualization, interaction, and data mining mechanisms. For example, outlier detection and named entity disambiguation algorithms could be applied to highlight wrong annotations and mentions which are likely referring to the same entity to practitioners in their annotation process. Concerning the network creation step (**Q1**), I proposed to model historical documents as bipartite multivariate dynamic networks to have a good balance between expressiveness and *simplicity*, while satisfying *traceability* and *document reality* properties. Since documents are explicitly represented as nodes in this model, the network entities are directly traceable to the original sources, thus enabling to reflect on the annotations while following socio-historical analyses with the same tools.

Leveraging the proposed network model, I presented the ComBiNet system in chapter 4 as a proof-of-concept to address **Q2**. By proposing easy-to-use exploration, visual querying, and comparison interactions on a model encoding all different dimensions of the content of historical documents (roles, social structure, location, time, other attributes), social historians were able to 1) reflect on their annotation process and potentially detect errors, while 2) answering complex historical questions on the specificities and differences of individuals and groups of interest. Finally, I proposed PK-Clustering in chapter 5, a new method for clustering based on the prior knowledge of social scientists, the consensus of automatic algorithms, and interactions for exploration. This system gives a concrete proposition to address **Q3** by providing a good balance between user control and data mining automatic capabilities while maintaining *traceability*, *simplicity*, and *document reality* properties, through detailed reports of interactions leading to the clustering, simple interactions mechanisms, and the use of bipartite multivariate dynamic networks as a data model. These two systems demonstrate that VA tools can support social historians in their overall workflow, and increase the traceability and control of the process while leveraging complex representations and algorithmic power. While ComBiNet and PK-Clustering have several limitations that I discuss in §6.2, I believe they lead the way towards better integration of VA tools to support social historians in their overall workflow, with a better level of simplicity and usability than state-of-the-art systems. Below, I discuss the perspective of new VA tools for HSNA, social history, and more globally the future of Digital Humanities in §6.3.

6.2 Discussion

I discuss in this section different limitations of my work:

Network modeling. I proposed with my collaborators to model historical documents using bipartite multivariate dynamic networks, as it allows to satisfy *traceability*, *simplicity*, and *document reality* properties. However, this type of modeling has some limitations in 1) the type of sources it can model, 2) how persons are represented in the network, and 3) how uncertainty is managed. We elaborated this model from collaborations with several social historians who study semi-structured documents, such as marriage acts, birth certificates, business contracts, construction documents, census, and migration forms. These types of documents have a consistent structure and mention people in a restricted number of relationships (spouses and witnesses for marriages, parents and children for birth certificates, etc.) that can be encoded as roles in a consistent manner. However, other types of textual documents can be leveraged by historians, which can be less structured or without any predefined structure at all. One example is correspondence letters, which is a type of document often studied in history [72, 195]. The content of letters is more verbose and varies from one to another, making the process of defining a set of relationships to encode more difficult. Bipartite multivariate dynamic networks would therefore not necessarily be an efficient model to encode this type of data, and other network models may be a better fit (such as directed networks). Moreover, in the proposed model, if documents are concretely represented as one type of node, person nodes constitute a construction made of the integration of several mentions of the same person from different sources. Historians, therefore, have to follow a named-entity-recognition and disambiguation process to give identifiers to the different persons mentioned in the documents and merge the information from several sources into one node. In one of my discussions with a historian, she mentioned that, in this model, "document nodes can be considered as *Emics* and person nodes *Etics* concepts" [107]. Historians hence have to make decisions, especially when there is ambiguity in the identity of persons, and when potentially contradicting information is written on the same individuals (concerning age, origin, profession, etc.). This process raises a problem that is widespread in quantitative social history, but also in most empirical science, which is the handling of *ambiguity* and *uncertainty*. Practitioners typically dismiss the uncertainty inherent to most textual data when constructing networks and encoding specific entities, thus removing it in the making of final conclusions. This is particularly true in history, where many mentions are ambiguous and not always precise [66]. Some work has been done on the handling of uncertainty in network models [3] and visualizations [206], but on simple network structures and not in the specific context of social history.

Temporality. The time is key information for historians, as they want to contextualize the phenomena they study in a period, relative to other events. This is why we encode time in our suggested model of bipartite multivariate dynamic networks through the time mentioned in historical documents, so historians can explore and analyze this dimension of their data. However, dynamic graphs are complex to visualize and analyze. In ComBiNet, if users can explore the dynamic aspect of the data through time distribution, overlays, and dynamic filters, it currently does not propose a layout unfolding the time structure. It may therefore be harder to detect time-related patterns compared to topological and geolocated ones (even if possible with interactions). PK-Clustering visualize the data through a static or dynamic layout. However, the current prototype considers only static clustering, which can be seen as a simplification of the real-world groups which can often evolve with time [197]. Indeed, persons can often meet new people, change affiliations, or move places, leading groups to merge, split, and disappear. PK-Clustering is already a complex process for static clustering but could be extended to the building of dynamic groups with the use of time-dependent prior knowledge and dynamic graph clustering algorithms.

HSNA and Social History. HSNA is now a widely used method in quantitative history to study relational structures and phenomena of the past [129, 183, 242]. The formalisms and tools proposed in this thesis aim at improving the workflow of historians following this type of method. Yet, historians usually have heterogeneous and various documents when they are researching an area and era of interest, and usually apply different methods at the same time to make their historical conclusions [179, 183]. The core of their work consists in extracting knowledge from rigorous inspection and cross-referencing of historical documents. If providing VA tools for their HSNA analysis from start to finish is useful to them, other types of analysis methods should also be implemented in their work environments to give them a larger set of options to make socio-historical conclusions. This includes methods like text analysis, correlations, and statistical testing [147]. History is also often considered a qualitative process, meaning that historians often make conclusions and hypotheses based on the reading of other sources and the qualitative analysis of their documents. VA tools that aim to encompass the whole historical workflow should be able to support this type of analysis, for example by managing textual annotation management on digital documents, similar to Jigsaw's feature for intelligence analysis [220]. Some quantitative methods can also let users express some of their qualitative knowledge to influence the results. For example, Bayesian statistics and semi-supervised machine learning methods are based on the expression of prior knowledge which will influence the computational results. Similarly, With PK-Clustering, historians can express their prior knowledge and use it as a start to find meaningful clusters, by seeing how the diversity of algorithms matches their qualitative knowledge of the data. VA tools for social history should therefore let users follow both qualitative and quantitative inspection of their documents, from data collection to final analysis, with combinations of several methods and prior knowledge expression.

Globality of the HSNA workflow. The key point of this thesis is to show that VA tools should support the overall HSNA workflow of historians. VA can be used to help them from data collection to their final analysis in the same environment, to ease back and forth between the steps, allowing easier exploration of different analysis goals, and better traceability/reproducibility for the overall analysis. By modeling historical documents into bipartite multivariate dynamic networks (see chapter 3), we represent the documents and their content as a network, allowing traceability between the network entities and the original documents. If historians find errors in the network, they can rapidly trace it back from which document the errors come from, and correct it either directly in the visual interface, or in their annotation software using the unique identifier of the document. This modeling choice is a first step towards better integration of the different steps into the same VA loop. Moreover, with ComBiNet, social scientists can apply filters to study specific visions of the network and follow multiple analysis paths on different dimensions of the data. ComBiNet, therefore, allows better integration of the annotation/encoding, modeling, and analysis/visualization steps, using the same interface. However, it does not allow complex network transformations (such as creating simple unipartite networks from projections) or creating new annotations in the texts of the documents. Historians still need to use ad-hoc methods for data collection and encoding and need to make their own scripts for complex network transformations. No tool currently exists which encompasses the whole workflow of historians, from the collection of documents to the final analysis.

6.3 Perspectives

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

Uncertainty models. As discussed in §6.2, historical sources are filled with ambiguity and imprecise information. Practitioners have to disambiguate mentions of homonyms to know if they refer to different persons or not. They also have to deal with potential surnames or plain errors in the writing of names. Similar problems arise for other entities' mentions such as locations since many places in the world have the same name or have changed names with time. With non-contemporary documents, location mentions can also have inconsistent resolution and refer to places with non-defined borders, such as "county of XX", or "kingdom of XX", which do not exist anymore. Moreover, historians can find contradictory information in several sources, for example concerning persons and events. When encoding their sources, practitioners hence have to make decisions on all this ambiguous information, by cross-referencing the documents, and using their common sense and intuition, to decide what seems the most probable choices. However, we could think about encoding schemes and data models which encapsulate the uncertainty inherent to historical data, to ground analysis results in a less biased and more realistic vision of the sources. I think this is a promising research direction, which has not stirred a lot of interest until now.

Dynamic Layouts and Clustering. As discussed in §6.2, temporality is one of the key dimensions of historical networks modeled as bipartite multivariate dynamic networks. Several layouts have been proposed to show the time aspect of dynamic graphs [14, 38] and bipartite dynamic graphs (which can be seen as dynamic hypergraphs), such as PAOHVIS [234], which is the layout PK-Clustering is based on. However, this type of layout does not allow us to see the attributes of persons and documents, nor the geolocation of entities. Moreover, most proposed layouts do not scale beyond approximately one hundred nodes. One way to solve this scalability problem is to aggregate the network, for example using dynamic clustering. Several dynamic clustering algorithms have been developed for dynamic networks, but they often struggle to take into account the complex dynamics of clusters, such as merging, splitting, or community drifting. Furthermore, no layout currently exists to visually display dynamic communities specifically for dynamic hypergraphs. More work can thus still be done in this space, to visualize the dynamic aspect of bipartite multivariate dynamic networks, and visualize dynamic communities. I started to develop a prototype to visualize the dynamic aspect of bipartite multivariate dynamic networks, with a focus on the document. Figure 6.2 shows the prototype. Each column corresponds to a timeslot and each square to a document. Persons mentioned in documents are displayed in the document nodes as smaller rectangles, colored by their roles.

For one person, all its mentions are linked through arcs inside one timestep, and splines through different timeslots. The splines are optimized to be of minimum size between each timestep. This representation should allow revealing the community structure of this type of data, by placing documents sharing many persons nearby. It also enables one to rapidly see properties associated with documents, for example by encoding the document nodes with color, as shown in Figure 6.2 (bottom).

This layout prototype could serve as a base to develop a process similar to PK-Clustering, but for creating meaningful dynamic groups, based on the consensus of dynamic clustering algorithms, prior knowledge of social historians, and exploration capabilities.

Machine Learning, automation, and agency. Machine learning went through rapid progress in the last 10 years, mainly due to the increase in data storage, computing power, and the rise of deep learning architectures. It has been applied to various tasks such as automatic driving, fraud detection, computer vision, and medical diagnostics. In the context of SNA, machine learning methods have been used to automatically extract knowledge through tasks such as node classification, clustering, and link prediction [161]. More broadly, it has also been used for historical document digitization [184]. If machine learning can give state-of-the-art accuracy on many of those tasks, it also poses issues with the explainability and reliability of the results in real-world applications. It can be particularly frustrating in the context of social history, as historians need to be able to understand and explain the structure of their networks, as discussed in chapter 5. Several methods and approaches now focus on trying to explain the outputs of these black-box algorithms to the end user [115]. Similarly, research is done on how to design interactive systems which leverage machine learning algorithms to guide and advise users, which are at the center of the decision-making process. This concept of utilizing artificial intelligence power to support human decision-making through interactive systems has been coined as "agency" [108] and "human-centered artificial intelligence" [216]. PK-clustering is based on the core idea that machine learning should support users while not removing their decision-making process, by providing automatic suggestions through clustering results, while letting social historians decide. ComBiNet could also be extended with machine learning suggestions, for example, to suggest social scientists recurring subgraphs in the data, that could be interesting to them. Over-represented subgraphs could be a query start that the users would refine through the easy-to-use visual query system. This idea of human-centered artificial intelligence could be applied not only to the analysis part but also to the data preparation workflow, for example in document transcription, named entity recognition, and disambiguation, all tasks that machine learning is efficient at.

A common workflow interface. Currently, most social scientists have to use a lot of different pieces of software, files, and ad-hoc processes to follow quantitative analyses. I provided two VA interfaces to help historians analyze their data and ease back and forth between the different steps of their analysis. Both interfaces use the same data format to lower the time cost to switch between them. However, historians still have to collect and annotate/encode their data manually with ad-hoc methods and may have to convert their data to various formats when using several visual analysis tools. All these operations usually break the traceability and reproducibility of their analyses, and make their process tedious, especially since it often requires

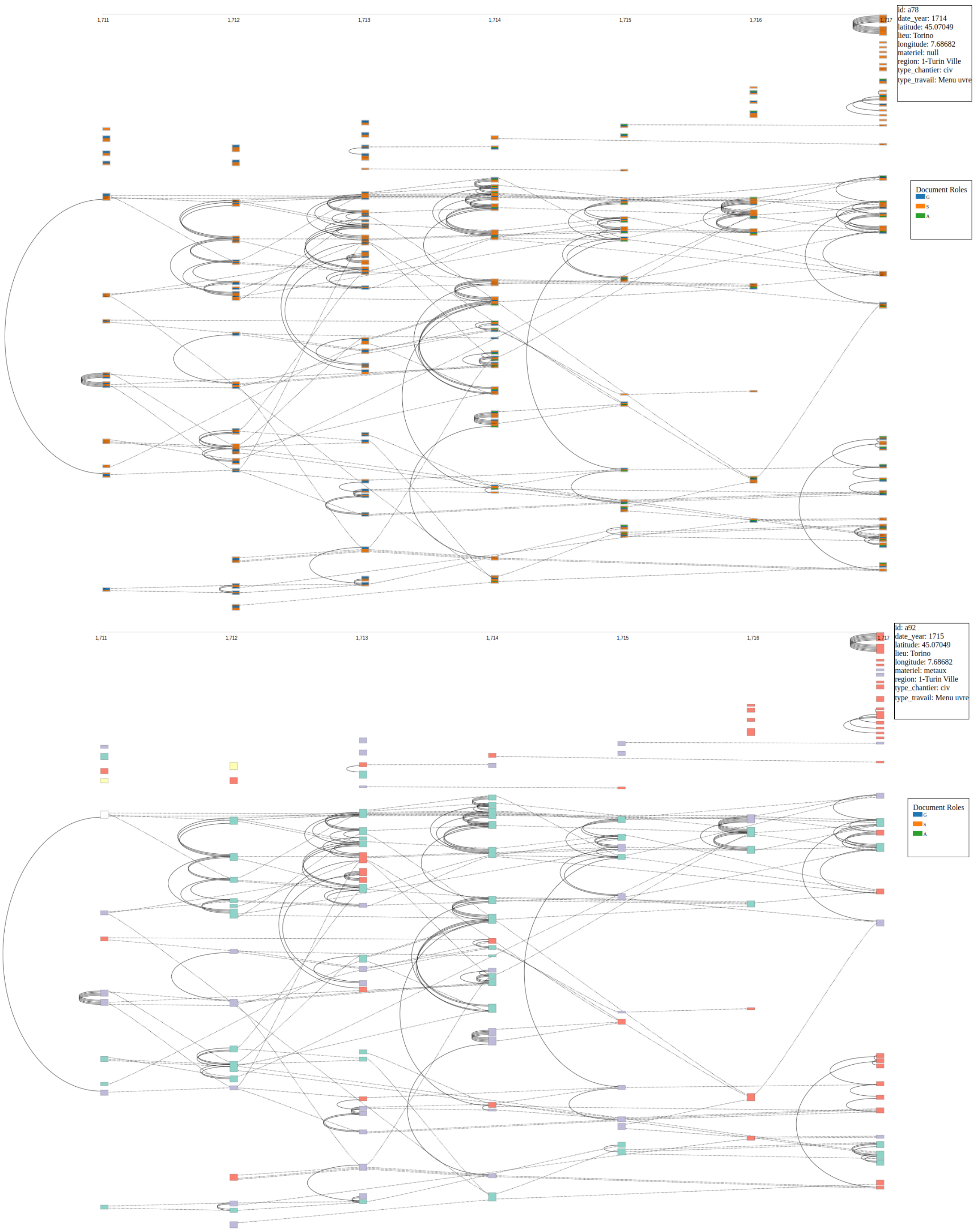


Figure 6.2 – Prototype of a document-centered dynamic layout for bipartite multivariate dynamic networks, showcased with construction contracts in Piedmont (see collaboration #1 in chapter 3). The layout can show the mentions of persons encoded with their roles (top) or the documents and their properties (bottom). Here, the *region* attribute is selected, hence coloring construction contracts depending on their locations.

the writing of conversion scripts, which they do not necessarily have the programming skills to do. I, therefore, argue there is a need for visual interfaces which integrate the whole workflow of social scientists, from the data collection to the formulation of high-level conclusions. If all the processes they do is integrated into the same visual environment, it would ease the flow of the analysis, increase the traceability and replicability of the actions and results, and allow them to take several exploration paths more easily. ComBiNet and PK-Clustering could for example be integrated into the same environment, with added possibilities of managing documents in the same place, applying annotations/encoding, and seeing in real time the creation of networks and transformations from the annotation process instead of having to do many back and forth. This constitutes an interesting research direction as it would allow social historians to collect, annotate, apply transform, analyze, and visualize their historical documents in the same environment, with easy-to-use interactions and artificial intelligence support.

6.4 Conclusion

To conclude, the goal of this thesis was to provide answers and directions towards the question of how VA, and more globally computer science, can help and support historians in the analyses they want to make. Towards this goal, I first formalized the current HSNA process from collaborations and discussions, defined three properties tools supporting this process should satisfy (*traceability*, *simplicity*, and *document reality*, and proposed two interfaces showcasing visualization and interactions mechanisms to support social historians in their workflow, leveraging historical documents modeled as bipartite multivariate dynamic networks. This network model explicitly represents the documents and the persons as nodes, thus enabling the representation of complex social relationships as they are mentioned without distortion and the tracing of networks' entities back to the documents, while keeping a good level of simplicity in regard to more complex models. ComBiNet allows users to explore this data model, and, therefore, reflect on their annotations, reveal specific facets of the data, and globally highlight and compare specific groups and behaviors to either detect erroneous patterns or answer socio-historical questions. PK-Clustering aims at integrating better the clustering task in the social historians' workflow, by providing a mixed-initiative approach for clustering on bipartite multivariate dynamic networks, leveraging the prior knowledge of practitioners, the consensus of automatic algorithms, and exploration capabilities. Both systems have been validated with real use cases, and aim at providing simple-to-use yet expressive tools, which let historians at the center of the analysis loop. Indeed, the use of quantitative methods in history and more globally the humanities has led to many expectations in the last 50 years, but also many disappointments due to usability and interpretability issues. More recently, with the rise of the popularity of machine learning, many propositions are made towards automatic inspection, extraction, and analysis of historical data based on artificial intelligence. Yet, similar criticisms of the beginnings of quantitative history emerged towards those methods, as they can easily lead to disembodied work in regard of the complexity of the content of historical documents.

Moreover, historians actually regard highly the inspection process of the sources. As one historian mentioned, "What many people promoting artificial intelligence to automatically read and inspect historical sources do not understand, is that this part is actually the most fun aspect of the historical work, and why many of us do it." Computer-supported tools for digital humanities should therefore support practitioners with interactive visualizations and quantitative-supported suggestions, instead of only providing automatic hard-to-interpret results.

It is the responsibility of the computer science community to focus on usability and users' control to produce tools adopted by the wide and diverse audience of social sciences. In contrast, social scientists still have to do the work for learning basic computer-science concepts and techniques such as file formats, data models, data encoding, and types (strings, integers), to be able to format their data at least basically to import them in computer-supported tools, and have more thorough discussions with computer scientists on their needs. To support the adoption of such tools by the humanities, both communities have to connect in meetups, allowing social scientists to share their workflows, research questions, and technical issues, and computer scientists to guide them on the preparation, managing, and computational analysis of their data. I think this collaborative vision constitutes a fertile ground for the production of computer-supported systems with low entry barriers which place social scientists at the center of the analysis loop by providing data-supported suggestions through visualization and easy-to-use interaction mechanisms. Propositions of this work aim in this direction, which I think is where lies the most difficult yet the most relevant area of the future of digital humanities as a field. Indeed, even though many methods and tools already exist to follow deep and complex analyses of sociological data, these are mostly only used by practitioners who claim themselves as digital humanists and not by the majority of social scientists who often get frustrated when using computer-supported tools and often need the help of computer scientists on the side. I think the only way to provide tools that truly empower the wider humanities' audience, is by creating the most flexible, expressive, and usable tools possible, which needs continuous discussions between the two communities (computer scientists and social scientists).

Bibliography

[1] Interchange: The Promise of Digital History. *Journal of American History*, 95(2):452–491, September 2008. doi:10.2307/25095630. 19

[2] Moataz Abdelaal, Nathan D. Schiele, Katrin Angerbauer, Kuno Kurzhals, Michael Sedlmair, and Daniel Weiskopf. Comparative Evaluation of Bipartite, Node-Link, and Matrix-Based Network Representations. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–11, 2022. doi:10.1109/TVCG.2022.3209427. 29

[3] Eytan Adar and Christopher Re. Managing Uncertainty in Social Networks. *IEEE Data Eng. Bull.*, 30:15–22, 2007. 111

[4] Ruth Ahnert, Sebastian E. Ahnert, Catherine Nicole Coleman, and Scott B. Weingart. The Network Turn: Changing Perspectives in the Humanities. *Elements in Publishing and Book Culture*, December 2020. doi:10.1017/9781108866804. 20

[5] Michael C. Alexander and James A. Danowski. Analysis of an ancient network: Personal communication and the study of social structure in a past society. *Social Networks*, 12(4):313–335, December 1990. doi:10.1016/0378-8733(90)90013-Y. xiii, 25, 26, 27

[6] Mashael Alkadi, Vanessa Serrano, James Scott-Brown, Catherine Plaisant, Jean-Daniel Fekete, Uta Hinrichs, and Benjamin Bach. Understanding barriers to network exploration with visualization: A report from the trenches. *IEEE Transactions on Visualization and Computer Graphics*, 27(2), 2022. doi:10.1109/TVCG.2022.3209487. 2, 5, 6, 9, 12, 25, 30, 31, 33, 34, 36, 40, 42, 48, 109

[7] Taher Alzahrani and K. J. Horadam. Community Detection in Bipartite Networks: Algorithms and Case studies. In Jinhu Lü, Xinghuo Yu, Guanrong Chen, and Wenwu Yu, editors, *Complex Systems and Networks: Dynamics, Controls and Applications*, Understanding Complex Systems, pages 25–50. Springer, Berlin, Heidelberg, 2016. doi:10.1007/978-3-662-47824-0_2. 87

[8] Keith Andrews, Martin Wohlfahrt, and Gerhard Wurzinger. Visual Graph Comparison. In *2009 13th International Conference Information Visualisation*, pages 62–67, July 2009. doi:10.1109/IV.2009.108. 55

[9] F. J. Anscombe. Graphs in Statistical Analysis. *The American Statistician*, 27(1):17–21, February 1973. doi:10.1080/00031305.1973.10478966. xiii, 14

[10] Thomas J. Archdeacon. *Correlation and Regression Analysis: A Historian's Guide*. Univ of Wisconsin Press, 1994. 36

[11] Mariona Coll Ardanuy, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel CS Wilson, and Barbara McGillivray. Living Machines: A study of atypical animacy, November 2020. arXiv:2005.11140, doi:10.48550/arXiv.2005.11140. 20

[12] David Auber, Daniel Archambault, Romain Bourqui, Maylis Delest, Jonathan Dubois, Antoine Lambert, Patrick Mary, Morgan Mathiaut, Guy Melançon, Bruno Pinaud, Benjamin Renoust, and Jason Vallet. TULIP 5. In Reda Alhajj and Jon Rokne, editors, *Encyclopedia of Social Network Analysis and Mining*, pages 1–28. Springer, August 2017. doi:10.1007/978-1-4614-7163-9_315-1. xiii, 16

[13] Benjamin Bach, Nathalie Henry Riche, Roland Fernandez, Emmanoulis Giannisakis, Bongshin Lee, and Jean-Daniel Fekete. NetworkCube: Bringing dynamic network visualizations to domain scientists. Posters of the Conference on Information Visualization (InfoVis), October 2015. 30

[14] Benjamin Bach, Emmanuel Pietriga, and Jean-Daniel Fekete. Visualizing dynamic networks with matrix cubes. In Matt Jones, Philippe A. Palanque, Albrecht Schmidt, and Tovi Grossman, editors, *CHI Conference on Human Factors in Computing Systems, CHI'14, Toronto, ON, Canada - April 26 - May 01, 2014*, pages 877–886. ACM, 2014. doi:10.1145/2556288.2557010. 113

[15] Juhee Bae, Tove Helldin, Maria Riveiro, Sławomir Nowaczyk, Mohamed-Rafik Bouguelia, and Göran Falkman. Interactive Clustering: A Comprehensive Review. *ACM Comput. Surv.*, 53(1):1:1–1:39, February 2020. doi:10.1145/3340960. 88

[16] Trevor J Barnes. Big data, little history. *Dialogues in Human Geography*, 3(3):297–302, November 2013. doi:10.1177/2043820613514323. 35

[17] Allen H. Barton. Survey Research and Macro-Methodology. *American Behavioral Scientist*, 12(2):1–9, November 1968. doi:10.1177/000276426801200201. 21

[18] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. In Eytan Adar, Matthew Hurst, Tim Finin, Natalie S. Glance, Nicolas Nicolov, and Belle L. Tseng, editors, *ICWSM*. The AAAI Press, 2009. 5, 29, 30, 53, 60, 84

[19] Sugato Basu, Ian Davidson, and Kiri Wagstaff. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. CRC Press, August 2008. 87

[20] Giuseppe Di Battista, Peter Eades, Roberto Tamassia, and Ioannis G. Tollis. *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall PTR, USA, 1st edition, 1998. 28

[21] Leilani Battle and Jeffrey Heer. Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau. *Computer Graphics Forum*, 38(3):145–159, 2019. doi:10.1111/cgf.13678. 56

[22] Michael Baur, Marc Benkert, Ulrik Brandes, Sabine Cornelsen, Marco Gaertler, Boris Köpf, Jürgen Lerner, and Dorothea Wagner. Visone Software for Visual Social Network Analysis. In Petra Mutzel, Michael Jünger, and Sebastian Leipert, editors, *Graph Drawing*, Lecture Notes in Computer Science, pages 463–464, Berlin, Heidelberg, 2002. Springer. doi:10.1007/3-540-45848-4_47. 29

[23] Laurent Beauguitte. Théorie des graphes et analyse de réseau en géographie : Histoire d'un lien faible (1950-1963). *PasserelleSHS*, 2022.

[24] Amine M. Bensaid, Lawrence O. Hall, James C. Bezdek, and Laurence P. Clarke. Partially supervised clustering for image segmentation. *Pattern Recognition*, 29(5):859–871, May 1996. doi:10.1016/0031-3203(95)00120-4. 87

[25] Jacques Bertin. *Sémiologie graphique: les diagrammes, les réseaux, les cartes*. Paris: Gauthier-Villars, 1967. xiii, 12, 13

[26] Anastasia Bezerianos, Fanny Chevalier, Pierre Dragicevic, Niklas Elmqvist, and Jean-Daniel Fekete. GraphDice: A System for Exploring Multivariate Social Networks. *Computer Graphics Forum*, 29(3):863–872, 2010. doi:10.1111/j.1467-8659.2009.01687.x. 61

[27] Marc Bloch and Jacques Le Goff. *Apologie pour l'histoire ou Métier d'historien*. Armand Colin, 1997. 2

[28] Christian Böhm and Claudia Plant. HISSCLU: A hierarchical density-based method for semi-supervised clustering. In *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '08, pages 440–451, New York, NY, USA, 2008. ACM. doi:10.1145/1353343.1353398. 87

[29] Georges-Pierre Bonneau, Hans-Christian Hege, Chris R. Johnson, Manuel M. Oliveira, Kristin Potter, Penny Rheingans, and Thomas Schultz. Overview and State-of-the-Art of Uncertainty Visualization. In Charles D. Hansen, Min Chen, Christopher R. Johnson, Arie E. Kaufman, and Hans Hagen, editors, *Scientific Visualization: Uncertainty, Multifield, Biomedical, and Scalable Visualization*, Mathematics and Visualization, pages 3–27. Springer, London, 2014. doi:10.1007/978-1-4471-6497-5_1. 89

[30] Stephen Borgatti. Social Network Analysis, Two-Mode Concepts in. *Computational Complexity: Theory, Techniques, and Applications*, January 2009. doi:10.1007/978-0-387-30440-3_491. 27, 45

[31] Christian Bors, John Wenskovitch, Michelle Dowling, Simon Attfield, Leilani Battle, Alex Endert, Olga Kulyk, and Robert S. Laramee. A Provenance Task Abstraction Framework. *IEEE Computer Graphics and Applications*, 39(6):46–60, November 2019. doi:10.1109/MCG.2019.2945720. 56

[32] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, December 2011. doi:10.1109/TVCG.2011.185. 60, 71

[33] Romain Boulet, Bertrand Jouve, Fabrice Rossi, and Nathalie Villa. Batch kernel SOM and related Laplacian methods for social network analysis. *Neurocomputing*, 71(7):1257–1273, March 2008. doi:10.1016/j.neucom.2007.12.026. xiii, 8, 25

[34] Pierre Bourdieu. Sur les rapports entre la sociologie et l'histoire en Allemagne et en France. *Actes de la Recherche en Sciences Sociales*, 106(1):108–122, 1995. doi:10.3406/arss.1995.3141. 17

[35] Paul Bradshaw. Data journalism. In *The Online Journalism Handbook*. Routledge, second edition, 2017. 15

[36] Ulrik Brandes. *Network Analysis: Methodological Foundations*. Springer Science & Business Media, February 2005. 5

[37] 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. doi:10.1109/TKDE.2007.190689. 40, 91

[38] Ulrik Brandes and Bobo Nick. Asymmetric Relations in Longitudinal Social Networks. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2283–2290, December 2011. doi:10.1109/TVCG.2011.169. 113

[39] Anne Burdick, Johanna Drucker, Peter Lunenfeld, Todd Presner, and Jeffrey Schnapp. *Digital_Humanities*. MIT Press, February 2016. 19

[40] Peter Burke. *History and Social Theory*. Polity, 2005. 11, 17

[41] Mitchell J. C. The Concept and Use of Social Networks. *Social Networks in Urban Situations*, 1969. 3, 21

[42] Steven P. Callahan, Juliana Freire, Emanuele Santos, Carlos E. Scheidegger, Cláudio T. Silva, and Huy T. Vo. VisTrails: Visualization meets data management. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data - SIGMOD '06*, page 745, Chicago, IL, USA, 2006. ACM Press. doi:10.1145/1142473.1142574. 41, 56

[43] Charles-Olivier Carbonell. *L'Historiographie*. FeniXX, January 1981. 17

[44] Stuart-K. Card, Jock-D. Mackinlay, and Ben Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers In, San Francisco, Calif, February 1999. 4, 12

[45] Marco Cavallo and Çağatay Demiralp. Clustrophile 2: Guided Visual Clustering Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):267–276, January 2019. doi:10.1109/TVCG.2018.2864477. 88

[46] Raphaël Charbey and Christophe Prieur. Stars, holes, or paths across your Facebook friends: A graphlet-based characterization of many networks. *Network Science*, 7(4):476–497, December 2019. doi:10.1017/nws.2019.20. 24, 54

[47] Duen Horng Chau, Christos Faloutsos, Hanghang Tong, Jason I. Hong, Brian Gallagher, and Tina Eliassi-Rad. GRAPHITE: A Visual Query System for Large Graphs. In *2008 IEEE International Conference on Data Mining Workshops*, pages 963–966, December 2008. doi:10.1109/ICDMW.2008.99. 55

[48] James Samuel Coleman. *Introduction to Mathematical Sociology*. Free Press of Glencoe / Collier-Macmillan, first edition edition, January 1964. 22

[49] Anna Collar, Fiona Coward, Tom Brughmans, and Barbara J. Mills. Networks in Archaeology: Phenomena, Abstraction, Representation. *J Archaeol Method Theory*, 22(1):1–32, March 2015. doi:10.1007/s10816-014-9235-6. 7

[50] TEI Consortium. TEI P5: Guidelines for electronic text encoding and interchange, February 2021. doi:10.5281/zenodo.4609855. 39

[51] Kristin Cook, Nick Cramer, David Israel, Michael Wolverton, Joe Bruce, Russ Burtner, and Alex Endert. Mixed-initiative visual analytics using task-driven recommendations. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 9–16, October 2015. doi:10.1109/VAST.2015.7347625. 88

[52] Ryan Cordell and David Smith. Viral texts: Mapping networks of reprinting in 19th-Century newspapers and magazines, 2017. 20

[53] Pascal Cristofoli. Aux sources des grands réseaux d'interactions. *Reseaux*, 152(6):21–58, 2008. 1, 7, 25, 34, 36, 38, 40, 44, 53

[54] Pascal Cristofoli. Principes et usages des dessins de réseaux en SHS. *La visualisation des données en histoire*, page 35, 2015. 2, 28, 60

[55] Pascal Cristofoli and Nicoletta Rolla. Temporalités à l'œuvre dans les chantiers du bâtiment. *Temporalités. Revue de sciences sociales et humaines*, (27), June 2018. doi:10.4000/temporalites.4456. xiv, 37, 43, 56, 61

[56] 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. doi:10.1016/j.socnet.2013.12.002. 43

[57] Alfred W. Crosby. *The Measure of Reality*. Cambridge University Press, Cambridge, reprint édition edition, March 1998. 4

[58] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. 55, 60

[59] Erick Cuenca, Arnaud Sallaberry, Dino Ienco, and Pascal Poncelet. VERTIGo: A Visual Platform for Querying and Exploring Large Multilayer Networks. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2021. doi:10.1109/TVCG.2021.3067820. 55, 82

[60] Zach Cutler, Kiran Gadhave, and Alexander Lex. Trrack: A Library for Provenance-Tracking in Web-Based Visualizations. In *2020 IEEE Visualization Conference (VIS)*, pages 116–120, October 2020. doi:10.1109/VIS47514.2020.00030. 68, 71

[61] Allison Davis, Burleigh Bradford Gardner, and Mary R. Gardner. *Deep South: A Social Anthropological Study of Caste and Class*. Univ of South Carolina Press, 2009. 45

[62] Mandeep K. Dhami, Ian K. Belton, and David R. Mandel. The "analysis of competing hypotheses" in intelligence analysis. *Applied Cognitive Psychology*, 33(6):1080–1090, 2019. doi:10.1002/acp.3550. 4

[63] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pages 269–274, New York, NY, USA, August 2001. Association for Computing Machinery. doi:10.1145/502512.502550. 92

[64] 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. 6, 31, 34, 38, 39

[65] 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. 37, 58

[66] Nicole Dufournaud. La recherche empirique en histoire à l'ère numérique. *Gazette des archives*, 240(4):397–407, 2015. doi:10.3406/gazar.2015.5321. 1, 111

[67] 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. xiii, 4, 6, 37, 42

[68] 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. doi:10.3166/dn.9.2.37-56. 39

[69] Nicole Dufournaud, Bernard Michon, Benjamin Bach, and Pascal Cristofoli. L'analyse des réseaux, une aide à penser : Réflexions sur les stratégies économique et sociale de

Marie Boucher, marchande à Nantes au XVII e siècle. In *Réseaux de Femmes, Femmes En Réseaux (XVIe-XXIe Siècles)*, pages 109–137. 2017.

[70] Cody Dunne, Nathalie Henry Riche, Bongshin Lee, Ronald Metoyer, and George Robertson. GraphTrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1663–1672, New York, NY, USA, May 2012. Association for Computing Machinery. doi:10.1145/2207676.2208293. 56

[71] Natural Earth. Free vector and raster map data. *Available: naturalearthdata. com*, 2015. 61

[72] Dan Edelstein, Paula Findlen, Giovanna Ceserani, Caroline Winterer, and Nicole Coleman. Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project. *The American Historical Review*, 122(2):400–424, April 2017. doi:10.1093/ahr/122.2.400. xiii, 19, 20, 33, 111

[73] 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. doi:10.1515/9781400841356.38. 22

[74] 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. doi:10.1086/502694. 43

[75] Michael Eve. Deux traditions d'analyse des reseaux sociaux. *Réseaux*, 115(5):183–212, 2002. 9, 24

[76] Wenfei Fan. Graph pattern matching revised for social network analysis. In *Proceedings of the 15th International Conference on Database Theory*, ICDT '12, pages 8–21, New York, NY, USA, March 2012. Association for Computing Machinery. doi:10.1145/2274576.2274578. 55

[77] Lucien Febvre. Vers une autre histoire. *Revue de Métaphysique et de Morale*, 54(3/4):225–247, 1949. 17

[78] Jean-Daniel Fekete, Danyel Fisher, Arnab Nandi, and Michael Sedlmair. *Progressive Data Analysis and Visualization*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, April 2019. doi:10.4230/DagRep.8.10.1. 67, 107

[79] Roderick Floud. *An Introduction to Quantitative Methods for Historians*. Routledge, London, September 2013. doi:10.4324/9781315019512. 36

[80] Professor Robert William Fogel. *Railroads and American Economic Growth: Essays in Econometric History*. The Johns Hopkins University Press, Baltimore, March 1970. 18

[81] Robert William Fogel. The Limits of Quantitative Methods in History. *The American Historical Review*, 80(2):329–350, 1975. doi:10.2307/1850498. 35

[82] Robert William Fogel and Stanley L Engerman. *Time on the Cross: Evidence and Methods, a Supplement*, volume 2. Little, Brown, 1974. 18

[83] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, February 2010. doi:10.1016/j.physrep.2009.11.002. 24, 87

[84] Linton Freeman. Visualizing Social Networks. *J. Soc. Struct.*, 2000. 2, 4, 5, 28

[85] Linton Freeman. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press, 2004. 1, 3, 11, 21, 22, 24, 40

[86] Manuel Freire, Catherine Plaisant, Ben Shneiderman, and Jen Golbeck. ManyNets: An interface for multiple network analysis and visualization. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, page 213, Atlanta, Georgia, USA, 2010. ACM Press. doi:10.1145/1753326.1753358. 55

[87] Michael Friendly. Visions and Re-Visions of Charles Joseph Minard. *Journal of Educational and Behavioral Statistics*, 27(1):31–51, March 2002. doi:10.3102/10769986027001031. 12

[88] 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. doi:10.1007/978-3-540-33037-0_2. 14

[89] Mohammad Ghoniem, Jean-Daniel 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, Austin, TX, USA, 2004. IEEE. doi:10.1109/INFVIS.2004.1. 29

[90] Carlo Ginzburg and Carlo Poni. La micro-histoire. *Le Débat*, 17(10):133, 1981. doi:10.3917/deba.017.0133. 3, 25, 36

[91] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, June 2002. doi:10.1073/pnas.122653799. 24, 87

[92] 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. 3

[93] Michael Gleicher. Considerations for visualizing comparison. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):413–423, 2018. doi:10.1109/TVCG.2017.2744199. 55

[94] Claudia Goldin. Cliometrics and the Nobel. *Journal of Economic Perspectives*, 9(2):191–208, June 1995. doi:10.1257/jep.9.2.191. 3

[95] Martin Grandjean. Social network analysis and visualization: Moreno's Sociograms revisited, 2015. xiii, 22

[96] Martin Grandjean. Analisi e visualizzazioni delle reti in storia. L'esempio della cooperazione intellettuale della Società delle Nazioni. *ME*, (2/2017), 2017. doi:10.14647/87204. 52

[97] Samuel Gratzl, Nils Gehlenborg, Alexander Lex, Hanspeter Pfister, and Marc Streit. Domino: Extracting, Comparing, and Manipulating Subsets Across Multiple Tabular Datasets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2023–2032, December 2014. doi:10.1109/TVCG.2014.2346260. 88

[98] Maurizio Gribaudi and Alain Blum. Des catégories aux liens individuels : l'analyse statistique de l'espace social. *Annales*, 45(6):1365–1402, 1990. doi:10.3406/ahess.1990.278914. 2

[99] Roger Guimerà, Marta Sales-Pardo, and Luís A. Nunes Amaral. Module identification in bipartite and directed networks. *Phys. Rev. E*, 76(3):036102, September 2007. doi:10.1103/PhysRevE.76.036102.

[100] Jo Guldi and David Armitage. *The History Manifesto*. Cambridge University Press, October 2014. 1, 20

[101] Aric Hagberg and Drew Conway. NetworkX: Network analysis with python. *URL: https://networkx. github. io*, 2020. 55, 84

[102] 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. doi:10.1080/1081602X.2014.892436. 25, 27, 43, 45

[103] 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. 1, 27

[104] J. Tuomas Harviainen and Bo-Christer Björk. Genealogy, GEDCOM, and popularity implications. *INF*, 37(3), October 2018. doi:10.23978/inf.76066. 27

[105] Mountaz Hascoët and Pierre Dragicevic. Interactive graph matching and visual comparison of graphs and clustered graphs. In Genny Tortora, Stefano Levialdi, and Maurizio Tucci, editors, *AVI '12*, pages 522–529. ACM, 2012. doi:10.1145/2254556.2254654. 55

[106] Loren Haskins and Kirk Jeffrey. *Understanding Quantitative History*. Wipf and Stock Publishers, March 2011. 16

[107] Thomas N. Headland, Kenneth L. Pike, and Marvin Harris, editors. *Emics and Etics: The Insider/Outsider Debate*. Emics and Etics: The Insider/Outsider Debate. Sage Publications, Inc, Thousand Oaks, CA, US, 1990. 42, 111

[108] Jeffrey Heer. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, 116(6):1844–1850, 2019. 114

[109] Jeffrey Heer and danah boyd. Vizster: Visualizing Online Social Networks. In John T. Stasko and Matthew O. Ward, editors, *IEEE Symposium on Information Visualization (InfoVis 2005), 23-25 October 2005, Minneapolis, MN, USA*, pages 32–39. IEEE Computer Society, 2005. doi:10.1109/INFVIS.2005.1532126. 62

[110] 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. doi:10.2307/1525715. 3

[111] 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. doi:10.1109/TVCG.2007.70582. xiii, 29

[112] Martin Hilbert and Priscila López. The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025):60–65, April 2011. doi:10.1126/science.1200970. 14

[113] 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. doi:10.1145/3447772. 43

[114] Danny Holten. Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):741–748, September 2006. doi:10.1109/TVCG.2006.147. 88

[115] Andreas Holzinger. From Machine Learning to Explainable AI. In *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pages 55–66, August 2018. doi:10.1109/DISA.2018.8490530. 114

[116] Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 159–166, New York, NY, USA, May 1999. Association for Computing Machinery. doi:10.1145/302979.303030. 85, 88

[117] Infovis SC policies FAQ. 102

[118] Piers J Ingram, Michael PH Stumpf, and Jaroslav Stark. Network motifs: Structure does not determine function. *BMC Genomics*, 7:108, May 2006. doi:10.1186/1471-2164-7-108. 54

[119] Petra Isenberg, Florian Heimerl, Steffen Koch, Tobias Isenberg, Panpan Xu, Charles D. Stolper, Michael Sedlmair, Jian Chen, Torsten Möller, and John Stasko. Vispubdata.org: A Metadata Collection About IEEE Visualization (VIS) Publications. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2199–2206, September 2017. doi:10.1109/TVCG.2016.2615308. 103

[120] Mikkel R. Jakobsen and Kasper Hornbæk. Evaluating a fisheye view of source code. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 377–386, New York, NY, USA, April 2006. Association for Computing Machinery. doi:10.1145/1124772.1124830. 107

[121] Radu Jianu, Adrian Rusu, Yifan Hu, and Douglas Taggart. How to Display Group Information on Node-Link Diagrams: An Evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 20(11):1530–1541, November 2014. doi:10.1109/TVCG.2014.2315995. 88

[122] J. David Johnson. UCINET: A software tool for network analysis. *Communication Education*, 36(1):92–94, January 1987. doi:10.1080/03634528709378647. 5, 29, 30

[123] J. David Johnson. UCINET: A software tool for network analysis. *Communication Education*, 36(1):92–94, January 1987. doi:10.1080/03634528709378647. 8

[124] 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. doi:10.1145/2682571.2797071. 5, 19

[125] 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. doi:10.1017/ahss.2019.90. 1, 35, 36, 41

[126] D.A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):59–78, January 2000. doi:10.1109/2945.841121. 107

[127] 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. doi:10.1007/978-3-540-70956-5_7. xiii, 5, 7, 15, 29

[128] Daniel Keim, Jörn Kohlhammer, Geoffrey Ellis, and Florian Mansmann. Mastering the information age: Solving problems with visual analytics. 2010. 103

[129] Florian Kerschbaumer, Linda Keyserlingk, Martin Stark, and Marten Düring. *The Power of Networks. Prospects of Historical Network Research*. Routledge, 2020. 1, 2, 11, 25, 112

[130] Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and cellular networks. *PLoS computational biology*, 5(5):e1000385, 2009. doi:10.1371/journal.pcbi.1000385. 81

[131] Jon Kleinberg. An Impossibility Theorem for Clustering. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. 85

[132] Dmitry A. Konovalov, Bruce Litow, and Nigel Bajema. Partition-distance via the assignment problem. *Bioinformatics*, 21(10):2463–2468, May 2005. doi:10.1093/bioinformatics/bti373. 92

[133] Elena V. Konstantinova and Vladimir A. Skorobogatov. Application of hypergraph theory in chemistry. *Discrete Mathematics*, 235(1-3):365–383, May 2001. doi:10.1016/S0012-365X(00)00290-9. 81

[134] 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. doi:10.1109/21.278993. xiii, 28

[135] R. Kosara, F. Bendix, and H. Hauser. Parallel Sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):558–568, July 2006. doi:10.1109/TVCG.2006.76. 88

[136] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. doi:10.1002/nav.3800020109. 93

[137] Alexander Kumpf, Bianca Tost, Marlene Baumgart, Michael Riemer, Rüdiger Westermann, and Marc Rautenhaus. Visualizing Confidence in Cluster-Based Ensemble Weather Forecast Analyses. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):109–119, January 2018. doi:10.1109/TVCG.2017.2745178. 88

[138] Ernest Labrousse. *La Crise de l'économie Française à La Fin de l'Ancien Régime et Au Début de La Révolution*, volume 1. Presses Universitaires de France-PUF, 1990. 17

[139] David S. Landes and Charles Tilly. *History as Social Science. The Behavioral and Social Sciences Survey*. Prentice Hall, Inc, 1971. 18

[140] Charles-Victor Langlois and Charles Seignobos. *Introduction aux études historiques*. ENS Éditions, February 2014. 1, 17

[141] Katherine A. Larson. Thomas F. Tartaron, Maritime Networks in the Mycenaean World. New York: Cambridge University Press, 2013. *Comparative Studies in Society and History*, 56(4):1064–1065, October 2014. doi:10.1017/S0010417514000516. 1

[142] Matthieu Latapy, Clémence Magnien, and Nathalie Del Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31–48, January 2008. doi:10.1016/j.socnet.2007.04.006. 45, 62

[143] Emmanuel Lazega. *Réseaux sociaux et structures relationnelles*. Presses universitaires de France, Paris, 1998. 5, 24

[144] Levi Lelis and Jörg Sander. Semi-supervised Density-Based Clustering. In *2009 Ninth IEEE International Conference on Data Mining*, pages 842–847, December 2009. doi:10.1109/ICDM.2009.143. 87

[145] Claire Lemercier. Analyse de réseaux et histoire. *Revue d'histoire moderne contemporaine*, 522(2):88–112, 2005. 38

[146] Claire Lemercier. 12. Formal network methods in history: Why and how? In Georg Fertig, editor, *Social Networks, Political Institutions, and Rural Societies*, volume 11, pages 281–310. Brepols Publishers, Turnhout, January 2015. doi:10.1484/M.RURHE-EB.4.00198. 1, 6, 7, 12, 23, 25, 27, 31, 33, 41, 52, 109

[147] Claire Lemercier and Claire Zalc. *Quantitative Methods in the Humanities: An Introduction*. University of Virginia Press, March 2019. 2, 3, 6, 7, 18, 19, 20, 27, 34, 35, 36, 41, 53, 112

[148] Claire Lemercier and Claire Zalc. Back to the Sources: Practicing and Teaching Quantitative History in the 2020s. *Capitalism*, 2(2):473–508, 2021. doi:10.1353/cap.2021.0010. 6, 18, 33, 34, 35, 36, 41

[149] Bernard Lepetit. L'histoire quantitative : deux ou trois choses que je sais d'elle. *Histoire & Mesure*, 4(3):191–199, 1989. doi:10.3406/hism.1989.1355. 1, 35

[150] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, December 2014. doi:10.1109/TVCG.2014.2346248. 96

[151] 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. doi:10.1177/0363199005278726. 25

[152] Carola Lipp and Lothar Krempel. Petitions and the Social Context of Political Mobilization in the Revolution of 1848/49: A Microhistorical Actor-Centred Network Analysis. *Int Rev of Soc His*, 46(S9):151–169, December 2001. doi:10.1017/S0020859001000281. 45

[153] Sehi L'Yi, Bongkyung Ko, DongHwa Shin, Young-Joon Cho, Jaeyong Lee, Bohyoung Kim, and Jinwook Seo. XCluSim: A visual analytics tool for interactively comparing multiple clustering results of bioinformatics data. *BMC Bioinformatics*, 16(11):S5, August 2015. doi:10.1186/1471-2105-16-S11-S5. 88

[154] Alan M. MacEachren, Anthony Robinson, Susan Hopper, Steven Gardner, Robert Murray, Mark Gahegan, and Elisabeth Hetzler. Visualizing Geospatial Information Uncertainty:

What We Know and What We Need to Know. *Cartography and Geographic Information Science*, 32(3):139–160, January 2005. doi:10.1559/1523040054738936. 89

[155] Stephen Makonin, Daniel McVeigh, Wolfgang Stuerzlinger, Khoa Tran, and Fred Popowich. Mixed-Initiative for Big Data: The Intersection of Human + Visual Analytics + Prediction. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 1427–1436, January 2016. doi:10.1109/HICSS.2016.181. 82, 88

[156] Gribaudi Maurizio. *Espaces, Temporalités, Stratifications :. Exercices Méthodologiques Sur Les Réseaux Sociaux*. Editions de l'Ecole des Hautes Etudes en Sciences Sociales, Paris, January 2000. 24

[157] Philip Mayer. Migrancy and the Study of Africans in Towns. *American Anthropologist*, 64(3):576–592, 1962. 24

[158] 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. 43

[159] Michael J. McGuffin. Simple algorithms for network visualization: A tutorial. *Tsinghua Science and Technology*, 17(4):383–398, August 2012. doi:10.1109/TST.2012.6297585. 29

[160] Pierre Mercklé and Claire Zalc. Peut-on modéliser la persécution ?: Apports et limites des approches quantifiées sur le terrain de la Shoah. *Annales. Histoire, Sciences Sociales*, 73(4):923–957, December 2018. doi:10.1017/ahss.2019.95. 1

[161] R. Michalski, P. Kazienko, and D. Krol. Predicting Social Network Measures Using Machine Learning Approach. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1056–1059, Istanbul, August 2012. IEEE. doi:10.1109/ASONAM.2012.183. 114

[162] Justin J Miller. Graph Database Applications and Concepts with Neo4j. page 7, 2013. 53, 55, 71, 80

[163] Barbara J. Mills. Social Network Analysis in Archaeology. *Annu. Rev. Anthropol.*, 46(1):379–397, October 2017. doi:10.1146/annurev-anthro-102116-041423. 1

[164] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594):824–827, October 2002. doi:10.1126/science.298.5594.824. 23, 54

[165] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. Independently published, 2020. 86

[166] Megan Monroe, Krist Wongsuphasawat, Catherine Plaisant, Ben Shneiderman, Jeff Millstein, and Sigfried Gold. Exploring point and interval event patterns: Display methods and interactive visual query. *University of Maryland Technical Report*, 2012. 80

[167] 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. doi:10.1037/10648-000. xiii, 22, 28

[168] J. L. Moreno. Foundations of Sociometry: An Introduction. *Sociometry*, 4(1):15, February 1941. doi:10.2307/2785363. 21

[169] Zacarias Moutoukias. Buenos Aires, port between two oceans: Mobilities, networks, stratifications (2nd half of the 18th century). *e-Spania. Revue interdisciplinaire d'études hispaniques médiévales et modernes*, 25, 2016. 37, 58

[170] 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. doi:10.3406/ahess.1992.279084. 25

[171] Andrej Mrvar and Vladimir Batagelj. Analysis and visualization of large networks with program package Pajek. *Complex Adaptive Systems Modeling*, 4(1), April 2016. doi:10.1186/s40294-016-0017-8. 5, 8, 29, 30, 60, 84

[172] Mark Newman. *Networks*. Oxford university press, 2018. 23

[173] Rolla Nicoletta. Mobilité et conflits. Travailler sur les chantiers de construction piémontais dans la première moitié du XVIIIe siècle. In Andrea Caracausi and Marco Schnyder, editors, *Travail et Mobilité En Europe (XVIe-XIXe Siècles)*, Coll. Histoire et Civilisations. Presses universitaires du Septentrion, Villeneuve d'Ascq, 2018. 37

[174] 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. doi:10.1109/TVCG.2018.2865149. 29

[175] Gérard Noiriel. Naissance du métier d'historien. *Genèses. Sciences sociales et histoire*, 1(1):58–85, 1990. doi:10.3406/genes.1990.1014. 17

[176] Jeremi K. Ochab, Jan Škvrňák, and Michael Škvrňák. Detecting Ottokar II's 1248–1249 uprising and its instigators in co-witnessing networks. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 0(0):1–20, May 2022. doi:10.1080/01615440.2022.2065397. 45

[177] Juri Opitz, Leo Born, and Vivi Nastase. Induction of a Large-Scale Knowledge Graph from the Regesta Imperii. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 159–168, Santa Fe, New Mexico, August 2018. Association for Computational Linguistics. 7, 53

[178] Maryjane Osa. *Solidarity And Contention: Networks Of Polish Opposition*. Univ Of Minnesota Press, Minneapolis, July 2003. 1, 40

[179] 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. doi:10.1086/230190. xiii, 1, 2, 6, 25, 112

[180] Terence J. Parr and Russell W. Quong. ANTLR: A predicated-LL (k) parser generator. *Software: Practice and Experience*, 25(7):789–810, 1995. 71

[181] Pamela Paxton. Dollars and Sense: Convincing Students That They Can Learn and Want to Learn Statistics. *Teach Sociol*, 34(1):65–70, January 2006. doi:10.1177/0092055X0603400106. 8, 42

[182] 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. doi:10.1177/14738716211045007. 47

[183] Cindarella Sarah Maria Petz. *On Combining Network Research and Computational Methods on Historical Research Questions and Its Implications for the Digital Humanities*. PhD thesis, Technische Universität München, 2022. 2, 36, 112

[184] James P. Philips and Nasseh Tabrizi. Historical Document Processing: Historical Document Processing: A Survey of Techniques, Tools, and Trends, September 2020. arXiv:2002.06300, doi:10.48550/arXiv.2002.06300. 114

[185] Robert Pienta, Fred Hohman, Alex Endert, Acar Tamersoy, Kevin Roundy, Chris Gates, Shamkant Navathe, and Duen Horng Chau. VIGOR: Interactive Visual Exploration of Graph Query Results. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):215–225, January 2018. doi:10.1109/TVCG.2017.2744898. 55

[186] Alexis Pister, Paolo Buono, Jean-Daniel Fekete, Catherine Plaisant, and Paola Valdivia. Integrating Prior Knowledge in Mixed-Initiative Social Network Clustering. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1775–1785, February 2021. doi:10.1109/TVCG.2020.3030347. 9, 30

[187] Alexis Pister, Nicole Dufournaud, Pascal Cristofoli, Christophe Prieur, and Jean-Daniel Fekete. From Historical Documents To Social Network Visualization: Potential Pitfalls and Network Modeling. In *7th Workshop on Visualization for the Digital Humanities (VIS4DH)*, 2022. 34

[188] Alexis Pister, Christophe Prieur, and Jean-Daniel Fekete. Visual Queries on Bipartite Multivariate Dynamic Social Networks. The Eurographics Association, 2022. doi:10.2312/evp.20221115. 52

[189] Antoine Prost. *Douze Leçons sur l'histoire*. Média Diffusion, April 2014. 2, 11, 16, 17

[190] Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, January 2007. doi:10.1093/bioinformatics/btl301. 23

[191] Eric Ragan, Endert Alex, Jibonananda Sanyal, and Jian Chen. Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes. *IEEE Transactions on Visualization and Computer Graphics*, 22(1), January 2016. doi:10.1109/TVCG.2015.2467551. 56

[192] Ramana Rao and Stuart K. Card. The table lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '94, pages 318–322, New York, NY, USA, 1994. Association for Computing Machinery. doi:10.1145/191666.191776. 107

[193] Donghao Ren, Bongshin Lee, and Matthew Brehmer. Charticulator: Interactive Construction of Bespoke Chart Layouts. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):789–799, January 2019. doi:10.1109/TVCG.2018.2865158. 56

[194] Pedro Ribeiro and Fernando Silva. Discovering Colored Network Motifs. In Pierluigi Contucci, Ronaldo Menezes, Andrea Omicini, and Julia Poncela-Casasnovas, editors, *Complex Networks V*, Studies in Computational Intelligence, pages 107–118, Cham, 2014. Springer International Publishing. doi:10.1007/978-3-319-05401-8_11. 55

[195] Christian Rollinger. Ciceros supplicatio und aristokratische Konkurrenz im Senat der Späten Republik. *Klio*, 99(1):192–225, August 2017. doi:10.1515/klio-2017-0007. 111

[196] Christian Rollinger. Prolegomena. Problems and Perspectives of Historical Network Research and Ancient History. *Journal of Historical Network Research*, 4:1–35, May 2020. doi:10.25517/jhnr.v4i0.72. 7, 8, 11, 12, 30, 42, 109

[197] Giulio Rossetti and Rémy Cazabet. Community discovery in dynamic networks: A survey. *ACM Comput. Surv.*, 51(2), February 2018. doi:10.1145/3172867. 26, 87, 112

[198] Giulio Rossetti, Letizia Milli, and Rémy Cazabet. CDLIB: A python library to extract, compare and evaluate communities from complex networks. *Appl Netw Sci*, 4(1):52, July 2019. doi:10.1007/s41109-019-0165-9. 87

[199] Fabrice Rossi, Nathalie Vialaneix, and Florent Hautefeuille. Exploration of a large database of French notarial acts with social network methods. *Digital Medievalist*, 9:2013, July 2014. doi:doi:10.16995/dm.52. 52

[200] Juan A. Rubio-Mondejar and Josean Garrues-Irurzun. Women entrepreneurs and family networks in Andalusia (Spain) during the second industrial revolution. *Business History*, pages 1–22, May 2022. doi:10.1080/00076791.2022.2068524. 1

[201] 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. 37

[202] 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. doi:10.1075/pbns.183.08sai. 43

[203] Bahador Saket, Paolo Simonetto, and Stephen Kobourov. Group-level graph visualization taxonomy. In N. Elmqvist, M. Hlawitschka, and J. Kennedy, editors, *EuroVis - Short Papers*. The Eurographics Association, 2014. doi:10.2312/eurovisshort.20141162. 88

[204] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350, 2016. doi:10.1109/TVCG.2016.2599030. 15, 71

[205] Shrutika S. Sawant and Manoharan Prabukumar. A review on graph-based semi-supervised learning methods for hyperspectral image classification. *The Egyptian Journal of Remote Sensing and Space Science*, 2018. doi:10.1016/j.ejrs.2018.11.001. 88

[206] Christoph Schulz, Arlind Nocaj, Jochen Goertler, Oliver Deussen, Ulrik Brandes, and Daniel Weiskopf. Probabilistic Graph Layout for Uncertain Network Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):531–540, January 2017. doi:10.1109/TVCG.2016.2598919. 111

[207] John Scott. Social Network Analysis. *Sociology*, 22(1):109–127, February 1988. doi:10.1177/0038038588022001007. 11, 21, 22, 23, 30, 40

[208] Michael Sedlmair, Christoph Heinzl, Stefan Bruckner, Harald Piringer, and Torsten Möller. Visual Parameter Space Analysis: A Conceptual Framework. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2161–2170, December 2014. doi:10.1109/TVCG.2014.2346321. 92

[209] 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. xiii, 29, 30, 31, 39

[210] Rachel Shadoan and Chris Weaver. Visual Analysis of Higher-Order Conjunctive Relationships in Multidimensional Data Using a Hypergraph Query System. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2070–2079, December 2013. doi:10.1109/TVCG.2013.220. 55

[211] Termeh Shafie, David Schoch, Jimmy Mans, Corinne Hofman, and Ulrik Brandes. Hypergraph Representations: A Study of Carib Attacks on Colonial Forces, 1509-1700. *Journal of Historical Network Research*, pages 52–70 Pages, October 2017. doi:10.25517/JHNR.V1I1.6. 45, 52

[212] Changxing Shang, Shengzhong Feng, Zhongying Zhao, and Jianping Fan. Efficiently detecting overlapping communities using seeding and semi-supervised learning. *Int. J. Mach. Learn. & Cyber.*, 8(2):455–468, April 2017. doi:10.1007/s13042-015-0338-5. 87

[213] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, September 1996. doi:10.1109/VL.1996.545307. 14, 15

[214] Ben Shneiderman. Dynamic queries for visual information seeking. *IEEE Software*, 11(6):70–77, November 1994. doi:10.1109/52.329404. 64

[215] Ben Shneiderman. Inventing Discovery Tools: Combining Information Visualization with Data Mining. *Information Visualization*, 1(1):5–12, March 2002. doi:10.1057/palgrave.ivs.9500006. 5, 15, 85

[216] Ben Shneiderman. *Human-Centered AI*. Oxford University Press, January 2022. 114

[217] Georg Simmel. *Soziologie: Untersuchungen über die Formen der Vergesellschaftung*. Duncker & Humblot, Berlin, 7. aufl edition, 2013. 23

[218] Marc A. Smith, Ben Shneiderman, Natasa Milic-Frayling, Eduarda Mendes Rodrigues, Vladimir Barash, Cody Dunne, Tony Capone, Adam Perer, and Eric Gleave. Analyzing (social media) networks with NodeXL. In John M. Carroll, editor, *Proceedings of the Fourth International Conference on Communities and Technologies, C&T 2009, University Park, PA, USA, June 25-27, 2009*, pages 255–264. ACM, 2009. doi:10.1145/1556460.1556497. 5, 29, 30, 53, 60

[219] John Snow. On the Mode of Communication of Cholera. *Edinb Med J*, 1(7):668–670, January 1856. 12

[220] John T. Stasko, Carsten Görg, and Zhicheng Liu. Jigsaw: Supporting investigative analysis through interactive visualization. *Inf. Vis.*, 7(2):118–132, 2008. doi:10.1057/palgrave.ivs.9500180. 31, 46, 47, 112

[221] Chris Stolte, Diane Tang, and Pat Hanrahan. Polaris: A System for Query, Analysis, and Visualization of Multidimensional Relational Databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):14, 2002. doi:10.1109/2945.981851. 47

[222] Lawrence Stone. The Revival of Narrative: Reflections on a New Old History. *Past & Present*, (85):3–24, 1979. 18

[223] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002. doi:10.1162/153244303321897735. 88

[224] Shazia Tabassum, Fabiola S. F. Pereira, Sofia Fernandes, and João Gama. Social network analysis: An overview. *WIREs Data Mining and Knowledge Discovery*, 8(5):e1256, 2018. doi:10.1002/widm.1256. 1, 5, 23

[225] R Core Team. R language definition. *Vienna, Austria: R foundation for statistical computing*, 2000. 55, 84

[226] Melissa Terras. Quantifying digital humanities. *URL: https://www.ucl.ac.uk/infostudies/melissa-terras/DigitalHumanitiesInfographic.pdf*, 2011. 19

[227] J. Joshua Thomas and Kris Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, January 2006. doi:10.1109/MCG.2006.5. 15

[228] Charles Tilly. *Retrieving European Lives*. Reliving the Past. Olivier Zunz, 1984. 2, 11, 16, 34, 52

[229] Charles Tilly. Observations of Social Processes and Their Formal Representations. *Sociological Theory*, 22(4):595–602, 2004. doi:10.1111/j.0735-2751.2004.00235.x. 1, 11, 35

[230] Natkamon Tovanich, Alexis Pister, Gaelle Richer, Paola Valdivia, Christophe Prieur, Jean-Daniel Fekete, and Petra Isenberg. VAST 2020 Contest Challenge: GraphMatchMaker: Visual Analytics for Graph Comparison and Matching. *IEEE Computer Graphics and Applications*, pages 1–1, 2021. doi:10.1109/MCG.2021.3091955. 54, 56

[231] Francesca Trivellato. Is There a Future for Italian Microhistory in the Age of Global History? *California Italian Studies*, 2(1), 2011. doi:10.5070/C321009025. 19, 36

[232] John W. Tukey. The Future of Data Analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962. 12

[233] John W. Tukey. *Exploratory Data Analysis*. Pearson, Reading, Mass, January 1977. 5, 15

[234] Paola Valdivia, Paolo Buono, Catherine Plaisant, Nicole Dufournaud, and Jean-Daniel Fekete. Analyzing Dynamic Hypergraphs with Parallel Aggregated Ordered Hypergraph Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(1):1–13, January 2021. doi:10.1109/TVCG.2019.2933196. 29, 47, 82, 89, 96, 113

[235] Guido van Rossum. Python Reference Manual. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995. 55

[236] 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. 43

[237] Corinna Vehlow, Fabian Beck, and Daniel Weiskopf. The state of the art in visualizing group structures in graphs. In R. Borgo, F. Ganovelli, and I. Viola, editors, *Eurographics Conference on Visualization (EuroVis) - STARs*. The Eurographics Association, 2015. doi:10.2312/eurovisstar.20151110. 88

[238] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al. Constrained k-means clustering with background knowledge. In *Icml*, volume 1, pages 577–584, 2001. 87

[239] Emily Wall, Subhajit Das, Ravish Chawla, Bharath Kalidindi, Eli T. Brown, and Alex Endert. Podium: Ranking Data Using Mixed-Initiative Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):288–297, January 2018. doi:10.1109/TVCG.2017.2745078. 88

[240] Wenjia Wang. Some fundamental issues in ensemble methods. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 2243–2250, June 2008. doi:10.1109/IJCNN.2008.4634108. 88

[241] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, November 1994. 5, 23, 30

[242] Charles Wetherell. Historical Social Network Analysis. *Int Rev of Soc His*, 43(S6):125–144, December 1998. doi:10.1017/S0020859000115123. 1, 3, 11, 25, 34, 40, 43, 52, 112

[243] Robert Whaples. Where Is There Consensus Among American Economic Historians? The Results of a Survey on Forty Propositions. *The Journal of Economic History*, 55(1):139–154, March 1995. doi:10.1017/S0022050700040602. 18

[244] Douglas R. White and Ulla Johansen. *Network Analysis and Ethnographic Problems: Process Models of a Turkish Nomad Clan*. Lexington Books, 2005. 1

[245] Hadley Wickham. An introduction to ggplot: An implementation of the grammar of graphics in R. *Statistics*, pages 1–8, 2006. 15

[246] Leland Wilkinson, D. Wills, D. Rope, A. Norton, and R. Dubbs. *The Grammar Of Graphics*. Springer-Verlag New York Inc., New York, 2nd ed. 2005 édition edition, 2005. 14

[247] Ian Winchester. The Linkage of Historical Records by Man and Computer: Techniques and Problems. *Journal of Interdisciplinary History*, 1(1):107, 1970. doi:10.2307/202411. 76

[248] Alvin W. Wolfe. The rise of network thinking in anthropology. *Social Networks*, 1(1):53–64, January 1978. doi:10.1016/0378-8733(78)90012-6. 4

[249] William Wright, David Schroh, Pascale Proulx, Alex Skaburskis, and Brian Cort. The Sandbox for analysis: Concepts and methods. In *Proceedings of the SIGCHI Conference*

on Human Factors in Computing Systems, CHI '06, pages 801–810, New York, NY, USA, April 2006. Association for Computing Machinery. doi:10.1145/1124772.1124890. 88

[250] Kai Xu, Alvitta Ottley, Conny Walchshofer, Marc Streit, Remco Chang, and John Wenskovitch. Survey on the Analysis of User Interactions and Visualization Provenance. *Computer Graphics Forum*, 39(3):757–783, June 2020. doi:10.1111/cgf.14035. 41, 56

[251] Jaewon Yang, Julian McAuley, and Jure Leskovec. Community Detection in Networks with Node Attributes. In *2013 IEEE 13th International Conference on Data Mining*, pages 1151–1156, December 2013. doi:10.1109/ICDM.2013.167. 87

[252] Franciszek Zakrzewski. The 1932 population register, May 2020. 36

[253] Michelle X. Zhou. "Big picture": Mixed-initiative visual analytics of big data. In *Proceedings of the 6th International Symposium on Visual Information Communication and Interaction*, VINCI '13, page 120, New York, NY, USA, 2013. Association for Computing Machinery. doi:10.1145/2493102.2499786. 88

[254] Tao Zhou, Jie Ren, Matúš Medo, and Yi-Cheng Zhang. Bipartite network projection and personal recommendation. *Phys. Rev. E*, 76(4):046115, October 2007. doi:10.1103/PhysRevE.76.046115. 92