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 xx décembre 2022, par

Alexis PISTER

THESE DE DOCTORAT

NNT : 2020UPASA001

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

université
PARIS-SACLAY

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 comment l'analyse visuelle peut supporter les historiens dans leur processus d'analyse de réseaux sociaux, de la collecte de documents historiques jusqu'à la formulation de conclusions socio-historiques. L'analyse de réseaux sociaux historiques est une méthode permettant d'étudier les relations sociales au sein de groupes d'acteurs (familles, institutions, entreprises, etc.) pour comprendre leurs structures sous-jacentes tout en décrivant des comportements spécifiques. Les chercheurs en histoire sociale reconstruisent les relations du passé à partir du contenu de documents historiques, tel que des actes de mariage, formulaires de migration, ou des recensements. Utilisant des méthodes analytiques et de visualisation, les historiens peuvent décrire la structure de ces groupes et expliquer des comportements individuels à partir de motifs locaux. Cependant, l'inspection, l'encodage et la modélisation des sources pour obtenir un réseau finalisé provoquent souvent des erreurs, distorsions et des problèmes de traçabilité. Pour ces raisons, ainsi que des problèmes d'utilisabilité, les historiens ne sont pas toujours en position de faire des conclusions approfondies sur leur réseau à partir des systèmes de visualisation actuels. Je vise dans cette thèse à identifier comment l'analyse visuelle (la combinaison d'algorithmes statistiques intégrés à des interfaces graphiques à l'aide d'interaction) peut supporter les historiens dans leur processus, de la collecte des données jusqu'à l'analyse finale. Vers ce but, je formalise le processus d'une analyse de réseau historique en partant de collaborations avec des historiens, de l'acquisition des sources jusqu'à l'analyse visuelle, et pointe que les outils supportant ce processus devraient satisfaire des principes de traçabilité, simplicité et de réalité documentaire pour faciliter les va-et-vient entre les différentes étapes, avoir des outils faciles à utiliser, et à ne pas distordre le contenu des sources. Particulièrement, je propose de modéliser les sources historiques en réseaux sociaux bipartis multivariés dynamiques avec rôles pour satisfaire ces propriétés. Ce modèle représente concrètement les documents historiques, permettant aux utilisateurs d'encoder, corriger et analyser leurs données avec le même modèle et les mêmes outils. Je propose deux interfaces d'analyse visuelle pour manipuler, explorer et analyser ce type de données, avec un appui sur les principes de traçabilité, simplicité, et réalité documentaire. Je présente d'abord ComBiNet, qui permet une exploration visuelle à partir de la topologie, dynamique, localisation et attributs du réseau à l'aide de vues coordonnées, un système de requêtes visuelles, et de comparaisons. En trouvant des motifs facilement et en les comparant, les historiens peuvent trouver des erreurs dans leurs annotations tout en répondant à des questions historiques. Le second système, PK-Clustering, constitue une proposition concrète pour améliorer l'utilisabilité et l'efficacité des mécanismes de clustering dans les systèmes de visualisation de réseaux sociaux. L'interface permet de créer des regroupements pertinent à partir de la connaissance à priori, le consensus algorithmique et l'exploration du réseau dans un cadre d'initiative mixte. Les deux systèmes ont été conçu à partir des besoins et de retours continus d'historiens, et visent à augmenter la traçabilité, simplicité, et la vision réelle des sources dans l'analyse de réseaux historiques. Je conclus sur des discussions sur la fusion des deux systèmes et plus globalement sur la convergence vers une meilleure intégration des outils d'analyse visuelle sur le processus global des historiens. De tels systèmes avec une attention les propriétés de traçabilité, simplicité, et réalité documentaire peuvent limiter l'introduction de biais et abaisser les exigences pour l'utilisation de méthodes quantitatives, qui a toujours été une discussion controversée 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 how Visual Analytics can support historians in their social network analysis process, from the collection of historical documents to the formulation of high-level socio-historical conclusions. Historical Social Network Analysis is a method to study social relationships between groups of actors (families, institutions, companies, etc.) to understand their underlying structure while characterizing specific behaviors. Social historians are able to reconstruct relationships of the past using historical documents' content, such as marriage acts, migration forms, birth certificates, and censuses. Through visualization and analytical methods, they can describe the global structure of studied groups and explain individual behaviors through local network patterns. However, the inspection, encoding, correction, and modeling process of the historical documents leading to a finalized network is intricate and often results in inconsistencies, errors, distortions, simplifications, and traceability issues. For these reasons, social historians are not always able to make thorough historical conclusions with current analytical and visualization tools. I aim in this thesis to identify how visual analytics—the integration of data mining capabilities into visual interfaces with interaction—can support social historians in their process, from the collection of their data to the answer to high-level historical questions. Towards this goal, I formalize the workflow of historical network analysis in collaboration with social historians, from the acquisition of their sources to their final visual analysis, and point out that visual analytics 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. Particularly, I propose to model historical sources into bipartite multivariate dynamic social networks with roles to satisfy those properties. This modeling allows a concrete representation of historical documents, hence letting users encode, correct, and analyze their data with the same abstraction and tools. Leveraging this data model, I propose two interactive visual interfaces to manipulate, explore, and analyze this type of data with a focus on usability for social historians. First, I present ComBiNet, which allows an interactive exploration leveraging the structure, time, localization, and attributes of the data model with the help of coordinated views, a visual query system, and comparison mechanisms. Finding specific patterns easily and comparing them, social historians are able to find inconsistencies in their annotations and answer their high-level questions. The second system, PK-Clustering, is a concrete proposition to increase the usability and effectiveness of clustering mechanisms in social network visual analytics systems. It consists in a mixed-initiative clustering interface that let social scientists create meaningful clusters with the help of their prior knowledge, algorithmic consensus, and 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. I conclude with discussions on the potential merging of both systems and more globally on research directions towards better integration of visual analytics systems on the whole workflow of social historians. Such systems with a focus on 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.

Contents

List of Figures

List of Tables

3 Historical Social Network Process, Pitfalls, and Network Modeling

Contents

I describe in this chapter a formalization of the HSNA workflow followed by social historians, to shed light on their process and summarize recurring pitfalls to identify how VA could support them in this workflow. Most HSNA practitioners report on their findings concerning the network they constructed from their sources, but few highlight the process which led to these conclusions from the raw historical documents, even though they have to make several annotation, encoding, and modeling decisions that deeply influence the final analysis [5]. Specifically, social historians can model documents and their content through various network models which have been proposed in the literature. I discuss in depth this step as it impacts the annotation and analysis possibilities, and I give an answer to our first research question **Q1** by proposing to model this type of data with bipartite multivariate dynamic networks. This model satisfies *simplicity*, *document reality*, and *traceability* properties, which we define as critical for social history work from our joint collaborations with social historians and current critics of HSNA [57, 116, 118].

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 [154]. 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 [117]. In this chapter, after describing and characterizing the workflow of HSNA [203] 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 [190]. 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 [5]. 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 [118] and how to model the annotation in a network [41]. 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 [51]. On the contrary, too complex models are complicated to visualize and analyze, and historians do not always have the tools to create them properly. In this chapter, I answer **Q1** (how to model historical documents into analyzable networks with the right balance between expressivity and simplicity) by proposing to model historical documents as bipartite multivariate dynamic networks, where both persons and documents are modeled as nodes with attributes and the links represent both individuals' mentions in the documents and their social roles in the event witnessed by the documents (such as witness in a marriage act). While this model is simple enough for creation and inspection, it allows tracing back the entities of the network to the original sources for a continuous annotation process and still accurately models the social relationships mentioned in the documents. Historians can therefore use this model to simultaneously find errors and inconsistencies in their annotation process—allowing them easier back and forth between the annotation and analysis steps—while starting a first analysis and exploration of the data to answer their sociological questions. The traceability to the original sources also makes the communications of findings more replicable and transparent.

3.2 Related Work

Since 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 [191]. 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 [10, 66, 119], 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 [101], 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" [191]. 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 [117, 118]. Most historical work leverage several methods in the same study to support their claims through different qualitative and quantitative results [150]. 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 [76]—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) [193].

3.2.2 Historian Workflows

Many quantitative methods and formalisms are available for historians to inspect their sources in the aim of making historical claims. Several textbooks describe and explain to social scientists and students who do not have formal computer science training in what consist these methods (statistical regression, Chi-squared test, network analysis, etc.) and how to practically use those with software and programming language [64]. 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 [5]. 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 [101, 117]. Lemercier and al. give guidelines on how to encode information from historical documents to prevent introducing bias, by having a critical view of the documents [118]. They emphasize the importance of the input phase of research and advise copying the first documents by hand while characterizing them in the most exhaustive and factual way, without imposing categorization. This explorative step let historians familiarize themselves with the content of the document, leading to a better view of what to encode to answer their research questions and sometimes to the formulation of new hypotheses. For example, in their project on the social and geographical trajectories of Jews in Lubartów [211], a village in Poland, the team noted the mean of writing inside the register documents (pen, pencil, ink, etc.) they were inspecting. This information allowed them to conclude that the inscription "expelled" written in pencil was probably added during World War II by Germans to denote exported Jews in the extermination camps. When applying network analysis, historians often create simplistic networks which allow them to answer specific research questions, but often loose this type of information related to the documents. Cristofoli discusses the network modeling problem when following a network analysis and highlights the fact that the same historical documents can be modeled in different ways [41], which can result in mismatches between the network shape and the research questions. Dufournaud presents her quantitative and network workflow when studying the economical role of women during the 16th and 17th centuries in the city of Nantes, which she splits into three steps: data collection, data processing, and data analysis [54].

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 historic 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 [43, 140].** 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 constructions, 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 across the years. Finally, she is interested in the average social mobility of individuals across the years.
3. Analysis of **marriage acts at Buenos Aires in the 17–19th centuries [134, 168].** 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 witness is sometimes used to ask favors in exchange.
4. Socio-political analysis of **migration of ethnic Germans from communist Romania to West Germany in the 20th century (ongoing work) [52].** The corpus is made of administrative forms that mention persons requesting to migrate, along with the persons they want to join, and the administrative persons of the ministry in charge of the forms. The family members of the aspiring migrants are also mentioned in the forms, with their respective dates of birth. 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.

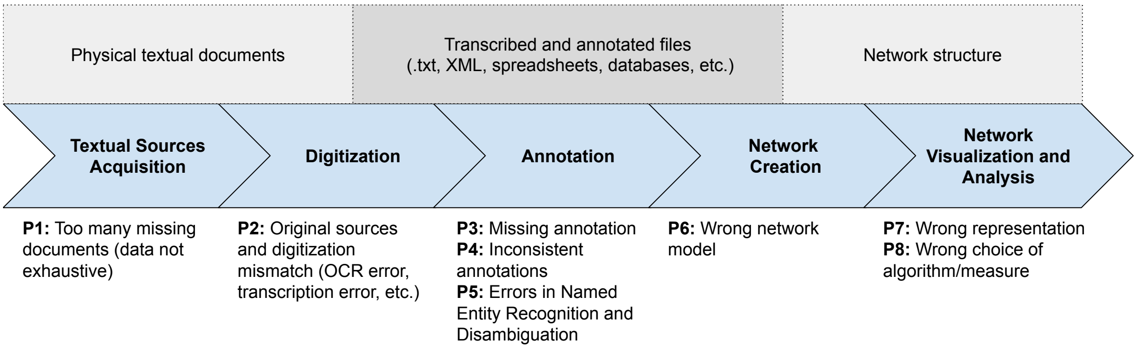

Physical textual documents
Transcribed and annotated files
(.txt, XML, spreadsheets, databases, etc.)
Network structure
Textual Sources Acquisition
Digitization
Annotation
Network Creation
Network Visualization and Analysis
P1: Too many missing documents (data not exhaustive)
P2: Original sources and digitization mismatch (OCR error, transcription error, etc.)
P3: Missing annotation
P4: Inconsistent annotations
P5: Errors in Named Entity Recognition and Disambiguation
P6: Wrong network model
P7: Wrong representation
P8: Wrong choice of algorithm/measure

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 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 [41, 51,]. 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. **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 [51], 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) [39] 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 [55,174]. 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 [5]. 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 [41]. 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 [70,203]. 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 [173]. 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 [28]. **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 [144] 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 [116]. 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 interface 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

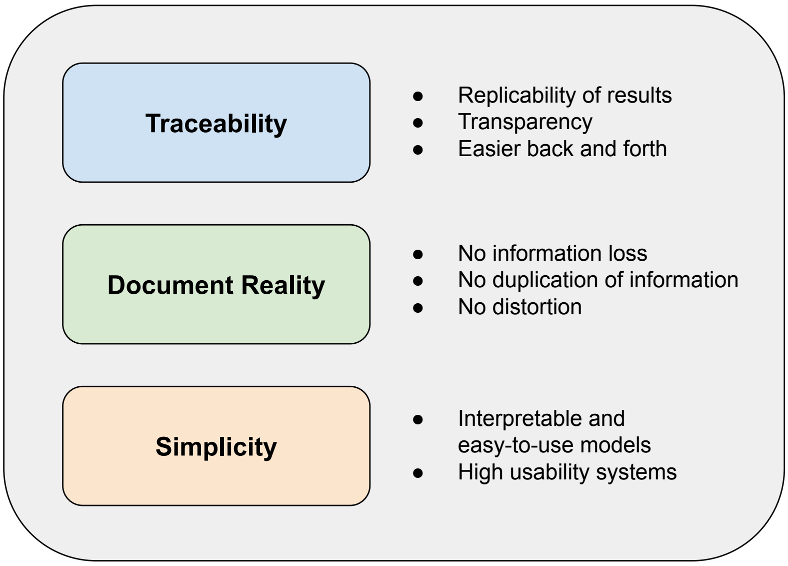


Figure 3.2 – Three properties essential to VA systems supporting the social historians workflow: *traceability*, *document reality*, and *simplicity*.

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 [32, 210]. 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 [101, 117]. Indeed, encoding and modeling the data with abstraction 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 [5] and often perform "soft SNA" [163] only due

[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 [101]. 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 [118] for specific examples.

[2] In anthropology, the terms *emics* and *etics* refer respectively to intrinsic phenomena related to observation and constructed categories and abstractions [90].

to usability problems and "Math anxiety" [148]. 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 [54]. 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 visual analytics tools should support social scientists in annotating and modeling their documents to make the HSNA process less linear by allowing easier back and forth between the annotation, modeling, and visualization steps. Network models satisfying *traceability*, *document reality*, 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.

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 [203]:** 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 [169]:** 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 [59]:** 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 [86] :** 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 [43]. 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 [125]:** 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 [44] and historians [198], 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 [96]:** 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.

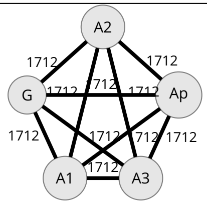
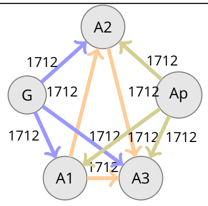
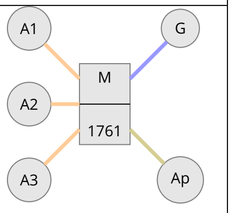
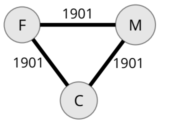
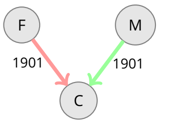
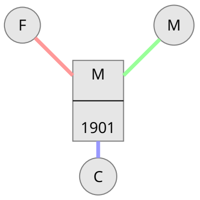
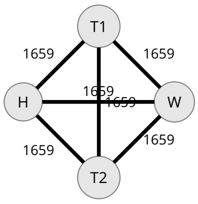
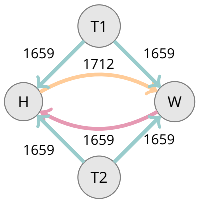
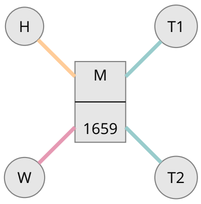
Original Document	Co-occurrence	Unipartite Multiplex	Bipartite
1712: Construction of a church in Torino. Associates: Bellotto G, Bello P.M, Bello G. Guarantor: Astrano G.A. Approbator: Corte A. Associate Guarantor Approbator			
Du dix-neuf fevrier mil huit cent quatre-vingt quatre, à six heures du soir. 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.

As shown by Cristofoli [41], 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 well simple relationships such as parenting ones as in example #2. However, it produces distortions for more complex relationships involving more than two persons, as in example #1 where people can either be mentioned as associates, guarantors, and approbators in the documents. Associates should probably be linked together with *associate* links, but the *guarantors* and *approbators* relationships are more complex to model. Approbators could be linked to the associates, the guarantors, or both. The three ways of modeling this type of relationship make sense but can lead to very different network shapes and analysis results. Historians thus have to decide on a transformation among several possibilities, which will probably distort the social reality of the relationships.

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 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 [41], 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 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 [?, 49, 121, 176] and the development of SNA bipartite measures and algorithms [22, 86, 113]. 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*, *witness*. This is a key aspect of our model since it clarifies the relationship between the persons within an event. In contrast, Jigsaw [183] 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 [184]. 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 (in the case of subgraph extraction or network modification interactions for example).

3.5 Applications

Several tools have been designed for visualizing dynamic bipartite networks that can also be considered dynamic hypergraphs [149, 196], 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 [183], 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

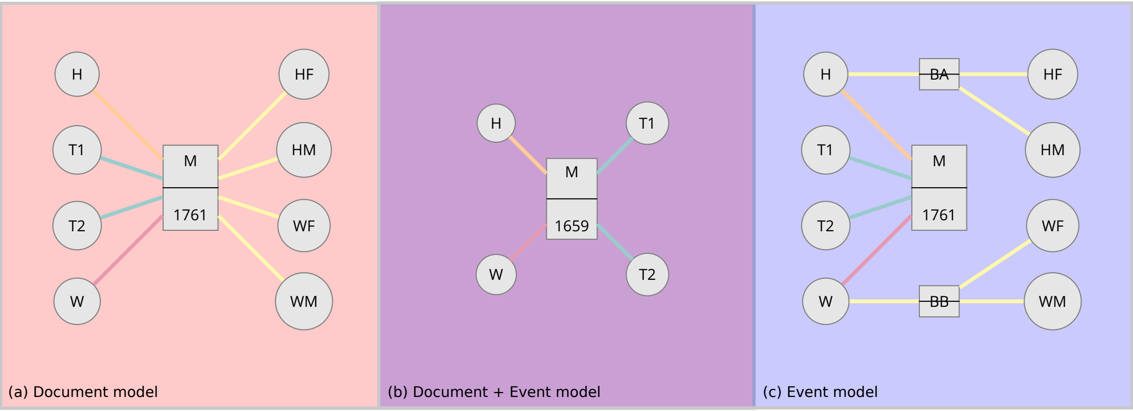


Figure 3.3 – bipartite multivariate dynamic network modeling for two cases of marriage acts of example #3. Some marriage acts mention the parents of the spouses, which is a relationship different than the marriage in itself. This case can be modeled using a document model (a) or an event model (c) by 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.

from going back to the original source, and supporting the social analyst in the annotation and modeling steps. We think visual analytics tools helping social scientists annotate and model their data with *document reality*, *traceability*, and *simplicity* principles in mind are essential to conducting socio-historical inquiries with limited friction, realistic training, and scientific transparency. Concerning the network modeling step, bipartite multivariate dynamic networks model well the majority of structured historical documents such as marriage acts, birth certificates, and business contracts as these documents refer to specific events (birth, marriage, transaction, etc). The document nodes, therefore, represent both the textual documents and the specific events. This dual representation works well for semi-structured documents but could be more limiting for other more literary documents. Moreover, structured documents can also provide information about other relationships not directly linked to the main event. For example, marriage acts sometimes refer to the place and date of birth of the spouses with the names of the parents. This information relates to the birth of the spouses and not the marriage specifically. In that case, social historians can either ignore this type of information in the annotation process or encode it with specific roles (*husband's father* and *wife's father* for example), thus turning the network into a model of the documents only, and not events. We show what would look like the resulting networks Figure 3.3for 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 [5]. 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 go back to the original source at any point of the workflow for easier corrections and replicability, and using models and methods simple and powerful enough for social scientists. Visual analytics software designed for HSNA should consider those principles to provide tools allowing to follow non-biased and reproducible analysis starting from the raw documents while supporting historians in going back and forth more easily between the annotation and analysis/visualization steps. I discussed the network modeling process in depth and claim that bipartite multivariate dynamic networks satisfies those three core principles, letting historians both wrangle their data and characterize sociological phenomena using a common model and visual representation, thus answering **Q1**. Using this model VA interfaces could help social scientists manage and analyze their data starting at the data acquisition and annotations steps instead 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 two next chapters.

Bibliography

[1] Interchange: The Promise of Digital History. *Journal of American History*, 95(2):452–491, September 2008. doi:10.2307/25095630. 33

[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, August 2022. arXiv:2208.04458. 42

[3] 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. 33

[4] 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. 9, 39, 41

[5] 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. 15, 21, 23, 26, 40, 43, 47, 48, 50, 53, 55, 63

[6] 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. 69

[7] F. J. Anscombe. Graphs in Statistical Analysis. *The American Statistician*, 27(1):17–21, February 1973. doi:10.1080/00031305.1973.10478966. 9, 28

[8] 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. 33

[9] 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. 9, 30

[10] Trevor J Barnes. Big data, little history. *Dialogues in Human Geography*, 3(3):297–302, November 2013. doi:10.1177/2043820613514323. 49

[11] Allen H. Barton. Survey Research and Macro-Methodology. *American Behavioral Scientist*, 12(2):1–9, November 1968. doi:10.1177/000276426801200201. 35

[12] 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. 20, 43, 67, 75, 98

[13] Sugato Basu, Ian Davidson, and Kiri Wagstaff. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC, first edition, 2008. 101

[14] 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. 42

[15] 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. 70

[16] 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. 43

[17] Jacques Bertin. *Sémiologie graphique: les diagrammes, les réseaux, les cartes*. Paris: Gauthier-Villars, 1967. 9, 26, 27

[18] A. Bezerianos, F. Chevalier, P. Dragicevic, N. Elmqvist, and J.d. Fekete. GraphDice: A System for Exploring Multivariate Social Networks. *Computer Graphics Forum*, 29(3):863–872, 2010. doi:10.1111/j.1467-8659.2009.01687.x. 75

[19] Marc Bloch. *Apologie Pour l'histoire*. A. Colin, 1949. 16

[20] 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. 101

[21] S.P. Borgatti, M. G. Everett, and L. C. Freeman. UCINET 6 for Windows: Software for Social Network Analysis. Harvard, MA, Analytic Technologies, 2002. 20

[22] 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. 40, 59

[23] 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. 70

[24] 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. 75, 85

[25] 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. 9, 22, 40

[26] 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. 31

[27] Paul Bradshaw. Data journalism. In *The Online Journalism Handbook*. Routledge, second edition, 2017. 29

[28] 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. 54

[29] Anne Burdick, Johanna Drucker, Peter Lunenfeld, Todd Presner, and Jeffrey Schnapp. *Digital_Humanities*. MIT Press, February 2016. 33

[30] Peter Burke. *History and Social Theory*. Polity, 2005. 31

[31] Mitchell J. C. The Concept and Use of Social Networks. *Social Networks in Urban Situations*, 1969. 16, 35

[32] 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. 55, 70

[33] Charles-Olivier Carbonell. *L'Historiographie*. FeniXX, January 1981. 31

[34] 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. 18, 26

[35] 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. 37, 38, 68

[36] 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. 69

[37] J. S. Coleman. Introduction to mathematical sociology. *Introduction to mathematical sociology.*, 1964. 36

[38] 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. 21

[39] TEI Consortium. TEI P5: Guidelines for electronic text encoding and interchange, February 2021. doi:10.5281/zenodo.4609855. 53

[40] Ryan Cordell and David Smith. Viral texts: Mapping networks of reprinting in 19th-Century newspapers and magazines, 2017. 33

[41] Pascal Cristofoli. Aux sources des grands réseaux d'interactions. *Reseaux*, 152(6):21–58, 2008. 15, 21, 39, 48, 50, 52, 54, 58, 59, 67

[42] Pascal Cristofoli. Principes et usages des dessins de réseaux en SHS. *La visualisation des données en histoire*, page 35, 2015. 16, 42, 75

[43] 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. 10, 51, 57, 70, 74

[44] 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. 57

[45] Alfred W. Crosby. *The Measure of Reality*. Cambridge University Press, Cambridge, reprint édition edition, March 1998. 18

[46] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. 69, 75

[47] 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. 69, 96

[48] 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. 82, 85

[49] 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. 59

[50] 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. 17

[51] 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. 21, 45, 48, 52, 53

[52] 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. 51, 72

[53] 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. 15

[54] 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. 9, 18, 21, 50, 56

[55] 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. 53

[56] 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. 70

[57] 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. 9, 33, 34, 47

[58] 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. 35

[59] 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. 57

[60] Michael Eve. Deux traditions d'analyse des reseaux sociaux. *Réseaux*, 115(5):183–212, 2002. 23, 37, 38

[61] 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. 69

[62] Lucien Febvre. VERS UNE AUTRE HISTOIRE. *Revue de Métaphysique et de Morale*, 54(3/4):225–247, 1949. 31

[63] 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. 81

[64] Roderick Floud. *An Introduction to Quantitative Methods for Historians*. Routledge, London, September 2013. doi:10.4324/9781315019512. 50

[65] Robert Fogel. *Railroads and American Economic Growth: Essays in Econometric History*. 1964. 32

[66] Robert William Fogel. The Limits of Quantitative Methods in History. *The American Historical Review*, 80(2):329–350, 1975. doi:10.2307/1850498. 49

[67] Robert William Fogel and Stanley L Engerman. *Time on the Cross: Evidence and Methods, a Supplement*, volume 2. Little, Brown, 1974. 32

[68] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, February 2010. doi:10.1016/j.physrep.2009.11.002. 37

[69] L. Freeman. Visualizing Social Networks. *J. Soc. Struct.*, 2000. 16, 18, 19, 41

[70] L.C. Freeman. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press, 2004. 15, 17, 25, 34, 35, 36, 37, 54

[71] Manuel Freire, Catherine Plaisant, Ben Shneiderman, and Jen Golbeck. ManyNets: An interface for multiple network analysis and visualization. In *CHI '10*, CHI '10, pages 213–222, New York, NY, USA, 2010. ACM. doi:10.1145/1753326.1753358. 69

[72] 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. 26

[73] 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. 28

[74] GEDCOM: The genealogy data standard. 41

[75] Mohammad Ghoniem, J.-D. Fekete, and Philippe Castagliola. A comparison of the readability of graphs using node-link and matrix-based representations. In *IEEE Symposium on Information Visualization*, pages 17–24. Ieee, 2004. 42

[76] Carlo Ginzburg and Carlo Poni. La micro-histoire. *Le Débat*, 17(10):133, 1981. doi:10.3917/deba.017.0133. 17, 38, 49

[77] 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. 37

[78] 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. 17

[79] Michael Gleicher. Considerations for visualizing comparison. *IEEE Trans. Vis. Comput. Graphics*, 24(1):413–423, 2018. doi:10.1109/TVCG.2017.2744199. 69

[80] Claudia Goldin. Cliometrics and the Nobel. *Journal of Economic Perspectives*, 9(2):191–208, June 1995. doi:10.1257/jep.9.2.191. 16

[81] Martin Grandjean. Social network analysis and visualization: Moreno's Sociograms revisited, 2015. 9, 36

[82] Martin Grandjean. Analisi e visualizzazioni delle reti in storia. L'esempio della cooperazione intellettuale della Societa&#768; delle Nazioni. *ME*, (2/2017), 2017. doi:10.14647/87204. 66

[83] 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. 15

[84] Jo Guldi and David Armitage. *The History Manifesto*. Cambridge University Press, October 2014. 15

[85] Aric Hagberg and Drew Conway. NetworkX: Network analysis with python. *URL: https://networkx. github. io*, 2020. 69, 98

[86] 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. 39, 41, 57, 59

[87] 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. 15, 41

[88] 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. 69

[89] Loren Haskins and Kirk Jeffrey. *Understanding Quantitative History*. Wipf and Stock Publishers, March 2011. 30

[90] 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. 55

[91] J. Heer and D. Boyd. Vizster: Visualizing online social networks. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 32–39, October 2005. doi:10.1109/INFVIS.2005.1532126. 76

[92] Jeffrey Heer. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, 116(6):1844–1850, 2019. 122

[93] 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. 16

[94] 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. 9, 42, 43

[95] 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. 28

[96] 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. 57

[97] Infovis SC policies FAQ. 115

[98] 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. 68

[99] J. David Johnson. UCINET: A software tool for network analysis. *Communication Education*, 36(1):92–94, January 1987. doi:10.1080/03634528709378647. 21, 43

[100] 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. 20, 33

[101] 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. 15, 49, 50, 55

[102] 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. 9, 20, 29

[103] Daniel A Keim. Visual Analytics. page 6. 43

[104] Florian Kerschbaumer, Linda Keyserlingk, Martin Stark, and Marten Düring. *The Power of Networks. Prospects of Historical Network Research*. January 2015. 15, 16, 25, 40

[105] Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and cellular networks. *PLoS computational biology*, 5(5):e1000385, 2009. 94

[106] Jon Kleinberg. An Impossibility Theorem for Clustering. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. 99

[107] 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. 94

[108] 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. 9, 42

[109] 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. 31

[110] David S. Landes and Charles Tilly. *History as Social Science. The Behavioral and Social Sciences Survey*. Prentice Hall, Inc, 1971. 32

[111] Charles-Victor Langlois and Charles Seignobos. *Introduction aux études historiques*. ENS Éditions, February 2014. 15, 30

[112] 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. 15

[113] 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. 59, 76

[114] Emmanuel Lazega. *Réseaux sociaux et structures relationnelles*. Presses universitaires de France, Paris, 1998. 19, 38

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

[116] 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. 15, 21, 25, 37, 39, 40, 41, 44, 47, 54, 66

[117] Claire Lemercier and Claire Zalc. *Quantitative Methods in the Humanities: An Introduction*. University of Virginia Press, March 2019. 16, 17, 21, 32, 33, 41, 48, 49, 50, 55, 67, 120

[118] 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. 21, 32, 47, 48, 49, 50, 55

[119] 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. 15, 49

[120] 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. 39

[121] 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. 59

[122] 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. 96

[123] Gribaudi Maurizio. *Espaces, Temporalités, Stratifications :. Exercices Méthodologiques Sur Les Réseaux Sociaux*. Editions de l'Ecole des Hautes Etudes en Sciences Sociales, Paris, January 2000. 37

[124] Philip Mayer. Migrancy and the Study of Africans in Towns. *American Anthropologist*, 64(3):576–592, 1962. 38

[125] 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. 57

[126] 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. 42

[127] 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. 15

[128] 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. 122

[129] 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. 37, 68

[130] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. Lulu.com, 2019. 100

[131] 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. 94

[132] 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. 9, 36, 41

[133] J. L. Moreno. Foundations of Sociometry: An Introduction. *Sociometry*, 4(1):15, February 1941. doi:10.2307/2785363. 35

[134] Zacarias Moutoukias. Buenos Aires, port between two oceans: Mobilities, networks, stratifications (2nd half of the 18th century). *E-SPANIA-REVUE ELECTRONIQUE D ETUDES HISPANIQUES MEDIEVALES*, 25, 2016. 51, 72

[135] 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. 39

[136] 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. 43, 75

[137] Natural earth. 75

[138] Neo4j graph data platform. 67, 69, 85, 94

[139] Mark Newman. *Networks*. Oxford university press, 2018. 36

[140] 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. 51

[141] 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. 42

[142] Gérard Noiriel. Naissance du métier d'historien. *Genèses. Sciences sociales et histoire*, 1(1):58–85, 1990. doi:10.3406/genes.1990.1014. 30

[143] 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. 21, 67

[144] Maryjane Osa. *Solidarity And Contention: Networks Of Polish Opposition*. Univ Of Minnesota Press, Minneapolis, first edition edition, July 2003. 15, 54

[145] 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. 9, 15, 19, 39, 40

[146] Pajek — Analysis and visualization of very large networks. 20, 21, 98

[147] Terence J. Parr and Russell W. Quong. ANTLR: A predicated-LL (k) parser generator. *Software: Practice and Experience*, 25(7):789–810, 1995. 85

[148] 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. 21, 56

[149] 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. 61

[150] 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. 16, 40, 49

[151] 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. 122

[152] 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. 69

[153] 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. 22, 43

[154] 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. 47

[155] 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. 66

[156] Antoine Prost. *Douze Leçons sur l'histoire*. Média Diffusion, April 2014. 16, 25, 30, 31

[157] Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, January 2007. doi:10.1093/bioinformatics/btl301. 37

[158] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. 69, 98

[159] 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. 70

[160] 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. 119

[161] 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. 70

[162] 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. 68

[163] 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. 21, 22, 25, 55

[164] Giulio Rossetti and Rémy Cazabet. Community discovery in dynamic networks: A survey. *ACM computing surveys (CSUR)*, 51(2):1–37, 2018. 40

[165] 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. 101

[166] 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. 66

[167] 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. 15

[168] 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. 51

[169] 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. 57

[170] 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. 102

[171] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. Vega-lite: A grammar of interactive graphics. *IEEE Trans. Vis. Comput. Graphics*, 23(1):341–350, 2016. 29, 85

[172] 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. 101

[173] John Scott. Social Network Analysis. *Sociology*, 22(1):109–127, February 1988. doi:10.1177/0038038588022001007. 25, 35, 36, 43, 54

[174] 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. 9, 42, 43, 44, 53

[175] 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. 69

[176] 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. 59, 66

[177] 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. 28, 29

[178] Ben Shneiderman. Dynamic queries for visual information seeking. *IEEE Softw.*, 11(6):70–77, November 1994. doi:10.1109/52.329404. 78

[179] 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. 18

[180] Georg Simmel. *Soziologie: Untersuchungen über die Formen der Vergesellschaftung*. Duncker & Humblot, Berlin, 7. aufl edition, 2013. 37

[181] 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. 20, 43, 67, 75

[182] John Snow. On the Mode of Communication of Cholera. *Edinb Med J*, 1(7):668–670, January 1856. 26

[183] 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. 45, 60, 61, 120

[184] 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. 61

[185] Lawrence Stone. The Revival of Narrative: Reflections on a New Old History. *Past & Present*, (85):3–24, 1979. 32

[186] 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. 102

[187] 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. 15, 19, 36

[188] Melissa Terras. Quantifying digital humanities. *UCL Centre for Digital Humanities*, 2011. 33

[189] J.J. Thomas and K.A. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, January 2006. doi:10.1109/MCG.2006.5. 29

[190] Charles Tilly. Retrieving european lives. 1984. 16, 25, 29, 48, 66

[191] 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. 25, 49

[192] 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. 68, 69

[193] 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. 32, 50

[194] John W. Tukey. The Future of Data Analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962. 26

[195] John W. Tukey. *Exploratory Data Analysis*. Pearson, Reading, Mass, 1er édition edition, January 1977. 19, 29

[196] Paola Valdivia, Paolo Buono, Catherine Plaisant, Nicole Dufournaud, and Jean-Daniel Fekete. Analyzing Dynamic Hypergraphs with Parallel Aggregated Ordered Hypergraph Visualization. *IEEE Trans. Visual. Comput. Graphics*, 27(1):1–13, January 2021. doi:10.1109/TVCG.2019.2933196. 42, 61, 96, 121

[197] Guido van Rossum. Python tutorial. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995. 69

[198] 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. 57

[199] 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. 102

[200] VisMaster: Visual analytics — Mastering the information age. 116

[201] 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. 101

[202] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, November 1994. 19, 37, 43

[203] Charles Wetherell. Historical Social Network Analysis. *Int Rev of Soc His*, 43(S6):125–144, December 1998. doi:10.1017/S0020859000115123. 15, 17, 25, 38, 39, 48, 54, 56, 66

[204] 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. 32

[205] Douglas White, Douglas R. White, and Ulla Johansen. *Network Analysis and Ethnographic Problems: Process Models of a Turkish Nomad Clan*. Lexington Books, 2005. 15

[206] Hadley Wickham and Maintainer Hadley Wickham. The ggplot package. *Google Scholar*, 2007. 29

[207] Leland Wilkinson. *The Grammar Of Graphics*. Springer, New York, 1993. 28

[208] 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. 90

[209] 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. 18

[210] 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. 55, 70

[211] Franciszek Zakrzewski. The 1932 population register, May 2020. 50

[212] 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. 101