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

Analyse Visuelle pour l'Analyse de Réseaux Sociaux Historiques

Visual Analytics for Historical Network Research

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

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

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

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

Alexis PISTER

Composition du jury

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

THESE DE DOCTORAT

NNT : 2020UPASA001

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

Titre: titre (en français)...

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

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

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

Title: titre (en anglais)...

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

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

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

Contents

3 - HSNA Process and Network Modeling

We describe in this chapter the HSNA workflow followed by social historians, to shed light on their process and summarize recurring pitfalls to identify how VA could help them in this process. Specifically, we discuss in depth the network modeling step, as the choice of the network model influence the overall process, especially the possibilities of the analysis. Most HSNA practitioners report on their findings concerning the network they constructed from their sources, but few highlight their process which led to these conclusions from the raw historical documents. Similarly, VA tools always focus on the analysis part, once the network have been constructed, without helping historians in the previous steps. However, the data collection, encoding, and transformation steps are crucial and can introduce lots of bias and distortion on the final data if not done correctly. This is especially true for social history where historical documents can lack structure and can be hard to parse, and where historical claims should be traceable to the original sources. We therefore describe the HSNA workflow split into 5 steps and characterize recurring pitfalls which can occur in each step. We also discuss in depth the network modeling step, as social historians can model their documents with various models which have an impact on the representation of the social relationships, traceability to the documents, and simplicity of usage.

This chapter is an updated version of an article presented at the VIS4DH workshop of the IEEE VIS: Visualization Visual Analytics Conference 2022, and published in IEEE Explore [112]. It was a collaboration with Nicole Dufournaud, Pascal Cristofoli, and my supervisors Christophe Prieur and Jean-Daniel Fekete. I participated in 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 generate many networks from the same documents/sources to visualize and analyze them. In this chapter, after describing and characterizing the workflow of Historical Social Network Analysis [149] from our collaborations with social historians, we explain why and how effective tools for supporting this process should model social networks in multiple steps to support three essential principles: *traceability*, connection to *reality*, and *simplicity*. These principles emerged from joint experiences as historians and computer scientists while collaborating on multiple projects.

Social historians' goal is to characterize socio-economic phenomena and their dynamics in a restricted period and place of interest and to see how individual people of that time lived through those changes. For this, they rely on historical documents such as conversational letters, censuses, and marriage acts. They usually extract qualitative and quantitative information from an identified corpus of documents, to then make conclusions on interesting socio-economic topics such as migrations, business dynamics, education, and kinship. For doing this, historians can apply HSNA methods, by modeling the social relationships between a set of entities—usually individuals—into a network. Historians therefore collect documents, annotate them, construct a network from the annotations that they finally analyze and visualize to validate or find new hypotheses. Unfortunately, the process is often linear, and it is common that, when visualizing their network, historians spot errors and inconsistencies in the annotations that they could have fixed if the process was iterative.

Moreover, historical documents are often complex and the annotation and modeling process can be done in many ways. Several network models have been proposed ranging from simple and specific ones like co-occurrence networks to more general and complex ones such as multilayer networks and knowledge graphs. Simple models allow answering specific questions and are easy to manipulate but are often too simplistic and may distort the information contained in the documents. Moreover, they often break the traceability from the analysis to the original documents, making the communication of findings less reproducible and the process of cleaning the annotations complicated. Indeed, errors and mismatches often occur in the annotation process, for example, due to entity disambiguation problems. On the contrary, too complex models are complicated to visualize and analyze, and historians do not always have the tools to create them properly. In this chapter, we propose to model historical datasets as bipartite multivariate dynamic networks, where both persons and documents are modeled as nodes with attributes. While this model is simple enough for creation and inspection, it allows tracing back the entities of the network to the original sources for a continuous annotation process and still accurately models the social relationships mentioned in the documents. Historians can therefore use this model to simultaneously find errors and inconsistencies in their annotation process—allowing them easier back and forth between the annotation and analysis steps—while starting a first analysis and exploration of the data to answer their sociological questions. The traceability to the original sources also makes the communications of findings more replicable and transparent.

3.2 . Related Work

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

The essence of the historical discipline is based on a critical approach of sources and involves considering peers' work. Traditional approaches to history often focus on the construction of a narrative, without necessarily adopting a systematic and problematized approach to the exploitation of original sources. Social history and the "Annales School" proposed a new approach to history, by trying to describe and characterize socio-economic phenomena of the past by rigorously extracting information from historical documents and making conclusions from them.

With similar aims, Glaser and Strauss developed the "Grounded Theory" [50] as a methodology for the humanities to build hypotheses and theories by solely studying and categorizing real-world observations, without starting from prior knowledge and predefined categories. Later on in the 1960s, quantitative methods started to be used in history, providing statistical and later computer-supported tools to aid historians in grounding their analysis in mathematical models and results. Unfortunately, the lack of methodology and understanding between the two worlds led to many criticisms by historians pointing to using wrong metrics, simplifying categories, and disconnections between the original documents and analysis [70, 82]. Quantitative history has been showed to be useful when used properly and when not focusing only on numbers, and several books have been published on how to efficiently use statistical methods such as summarizations, correlations, statistical distributions, statistical testing, time series etc. [66, 81]. Similarly, the use of network science for historical aims increased in recent years, and a lot or resources exist on how to use network methods and measures for historical research [72, 80].

However, little work has been done on describing and formalizing the process before the analysis part for a quantitative and network research workflow. Indeed, if it is central to know how to manipulate statistical and network concepts and methods when following this kind of methodology, it is as important if not even more to follow a correct and rigorous workflow to generate the data we plan to analyze beforehand. The process to generate a clean quantitative or network dataset from historical sources is difficult and requires several data acquisition, annotation, and cleaning steps. Social analysts are not always trained on how to do these steps effectivity, which can lead to errors, inconsistencies, and mismatches between the chosen data models and the historical questions [2]. Karila-Cohen and al. provide some advice on how to annotate historical documents with the aim of using quantitative methods [70] and prone that the annotation and analytical processes should not be dispatched between several persons, as both usually influence each other. Dufournaud describes her workflow in depth when studying the socio-economic status of women in France in the 16th and 17th centuries, which she splits into three steps: *data collection*, *data processing*, and *data analysis* [34]. She provides the tools and methodology she used to annotate her data, providing transparency on her historical analysis and methodological resources. Cristofoli discusses the network modeling problem when following an HSNA and highlights the fact that the same historical documents can be modeled in different ways [23]. Historians should be aware of this and choose a network model which fits their analytical

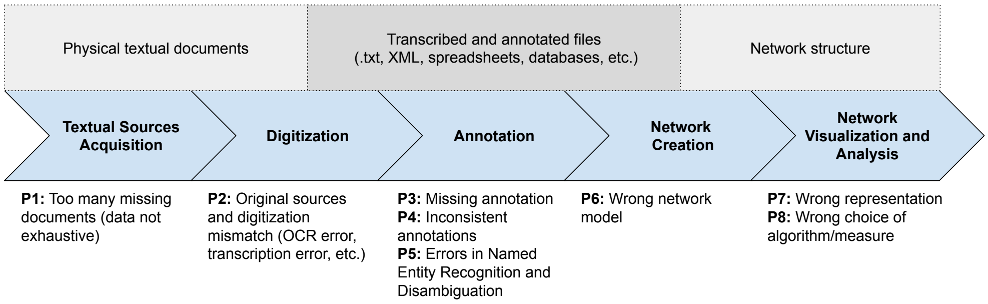


Figure 3.1 – HSNA workflow is split into five steps: textual sources acquisition, digitization, annotation, network creation, network visualization and analysis. We list potential pitfalls for each step.

goals.

3.3 . Historical Social Network Analysis Workflow

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

3.3.1 . Textual Sources Acquisition

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

3.3.2 . Digitization

Digitization consists in converting the sources into a digital format. This step can be skipped for the most recent periods where many documents have been produced digitally or can be scanned and well digitized through optical character recognition (OCR), allowing to tremendously ease the storage, indexation, and annotation of the documents. However, before mid 20th century, most historical primary sources are stored in archives in paper format and need human work to be digitized. **Mismatches between the original documents and the transcription can occur for old and recent documents (P2)**. However, if OCR tools are more and more efficient in English and highly used languages, historians can work with old documents written in old or extinguished languages and with atypical writings (e.g., Fraktur handwriting and typefaces for German in the early 20th century). Therefore, OCR tools are often unusable in social history and digitization remains an expensive and sometimes highly skilled process.

3.3.3 . Annotation

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

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

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

3.3.4 . Network Creation

Historians construct a network from the annotations of the documents. Usually, all persons mentioned are annotated and will be transformed into network nodes (vertices). Additional information such as their age, profession, and gender can be stored as node attributes. How the network's links are created is not as trivial and can vary from project to project [2]. The most straightforward approach is to create a link between every pair of persons mentioned in one document, thus forming a clique motif. This is a simplistic heuristic as social relationships can be quite complex, involving more than two persons who can have different roles in the relationship. The choice of the network model has a major impact on the future analysis and **may add bias if chosen loosely (P6).** More complex models have been proposed in the literature such as weighted, dynamic, bipartite, and layered networks.

3.3.5 . Network Analysis and Visualization

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

3.4 . Network modeling and analysis

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

3.4.1 . Network Models

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

- **Simple Networks [149]:** According to their research hypotheses, historians select and merge document information to build a specific relationship between individuals. They analyze this simple network structure with SNA tools and produce network indicators and node-link visualizations. It is often difficult to connect the results to the original sources.
- **Co-occurrence networks [123]:** Only the persons are represented as nodes, and two persons are connected with a link when they are mentioned in the same document (or section). This is a simple model and one of the first to have been used in SNA and HSNA. The major drawback of this model is that it does not take into account the diversity of social relationships, as every link is

identical. It can work well when only one type of social relationship is studied like a friendship network [95]. However, historical documents rarely mention only one type of relationship and this model is thereby very limiting for HSNA.

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

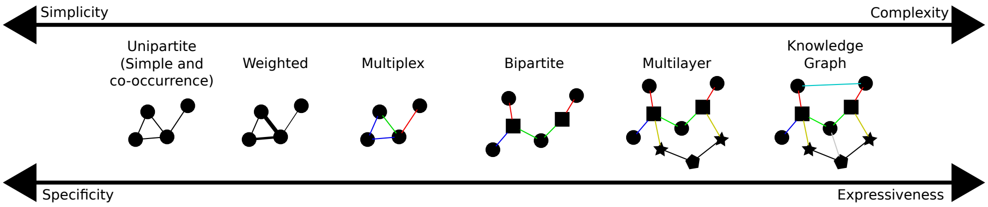


Figure 3.2 – Schematic representations of Different network models used for analyzing historical documents, ordered by complexity and expressiveness

We argue that historians should aim to model their networks simply enough to be manipulated by them, in a way that entities can be traced back to the sources, and expressive enough to model accurately the social reality of the documents—i.e., having those three properties: *simplicity*, *traceability*, and *reality*.

Currently, most digital historical projects use unipartite networks (simple, co-occurrence, and multiplex) that are simple and allow answering specific questions, but they do not capture all the complexity of the documents, and social scientists may miss important patterns. For example, modeling only co-occurrences of persons in documents remove the variety of social relationships these mentions can refer to. Moreover, since documents are not explicit in the unipartite model, it is hard to trace the network entities back to the sources: the traceability property is not satisfied. On the other side, multilayer networks and KG allow to model documents as entities and express complex relationships between various other entities they mention. These models can be very expressive but are challenging to use for historians, especially without guidelines; without *simplicity*, the *traceability* and *reality* properties can be hard to achieve. Moreover, they are difficult to visualize and analyze, especially for social scientists.

Figure 3.2 shows a schematic representation of the different network models, ranked on simplicity/complexity and specificity/expressiveness axis.

3.4.2 . Bipartite Multivariate Dynamic Social Network

Historical documents are well modeled by bipartite multivariate dynamic networks with roles, which have the following properties:

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

Links and Roles: A link models the mention of a person in a document. **Each link has a type corresponding to the role of the person in the document**. For a marriage act, the roles include *wife*, *husband*, *witness*. This is a key aspect of our model since it clarifies the relationship between the persons within an event.

In contrast, Jigsaw [136] does not consider the roles.

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

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

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

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

3.4.3 . Examples

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

1. Analysis of the social dynamics from **construction contracts in Italy in the 18$^{\text{th}}$ century [25, 101].** The corpus is made of contracts for different types of constructions in the Piedmont area in Italy. People are mentioned under three different roles: *Associates* who are in charge of the construction, *Guarantors* who bring financial guarantees, and *Approvers*, who vouch for the guarantors. Documents contain information about the building site, the type and materials of constructions, and the origin of the people.
2. Analysis of migrations from the **genealogy of a french family between the 17$^{\text{th}}$–20$^{\text{th}}$ centuries** [unpublished work]. The corpus is made of family trees referring to several document/event types: birth and death certificates, marriage acts, military records, and census reports. The roles are different for each event type and consist of *children, father, mother* for the birth events, *deceased* for the death event, *spouse* and *witnesses* for the marriages, and *family member*s for the census events.

3. Analysis of migrations from Spain to Argentina through the **marriage acts at Buenos Aires in the 17–19$^{\text{th}}$ centuries [96, 122].** The corpus is made of summaries of marriage records that mention the spouses and the witnesses of the wedding. The origin, date of birth, and parents' names are specified for both spouses.
4. Socio-political analysis of **migration of ethnic Germans from communist Romania to West Germany in the 20th century (ongoing work) [32].** The corpus is made of administrative forms that mention persons requesting to migrate, along with the persons they want to join, and the administrative persons of the ministry in charge of the forms. The family members of the aspiring migrants are also mentioned in the forms, with their respective dates of birth.

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

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

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

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

links in the co-occurrence model and in 9 links in the multiplex unipartite model while it is only stored once as a document node attribute in the bipartite model.

3.5 . Applications

Several tools have been designed for visualizing dynamic bipartite networks that can also be considered dynamic hypergraphs [108, 142], but few incorporate attributes. Moreover, the vast majority of visual analytics tools are solely focused on the analytical part of the data, meaning that the link between the original documents and the hypergraph abstraction is often broken. Social scientists therefore always have to do many back and forth between the visual analytics tools and their original documents and the annotation/modeling processes. More visual analytical tools should thus incorporate the textual documents in their data model similarly to Jigsaw [136], as it would allow tracing the entities of the network back to the original documents more easily. Mechanisms to clean/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 clean errors and inconsistencies in the annotations and propagate them in the visual analysis workflow. For example, the Vistorian [128] now lets users modify and clean their data in a table format if they see errors or inconsistencies.

3.6 . Discussion

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

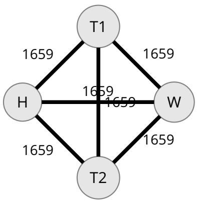
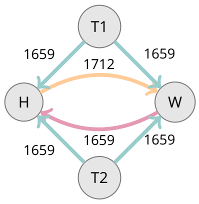
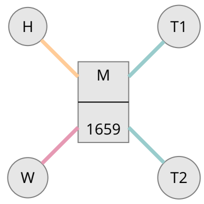
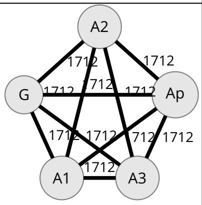
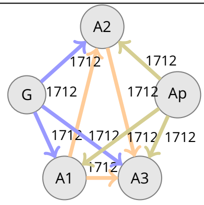
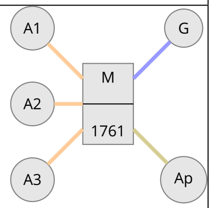
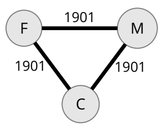
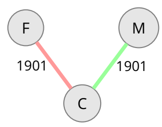
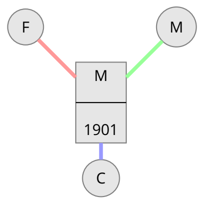
<table>
<tr><th>Original Document</th><th>Co-occurrence</th><th>Unipartite representation</th><th>Bipartite</th></tr>
<tr>
<td>20-4-1659 : Capitán Alonso MUÑOZ de GADEA , con Da. Francisca CABRAL LEAL de AYALA . Ts.: Agustín Gayoso , y Juan Guerrero. Al margen: "fue Oficial Real" , (f. 9v).
Husband Wife Witness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>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</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>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</td>
<td></td>
<td></td>
<td></td>
</tr>
</table>

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

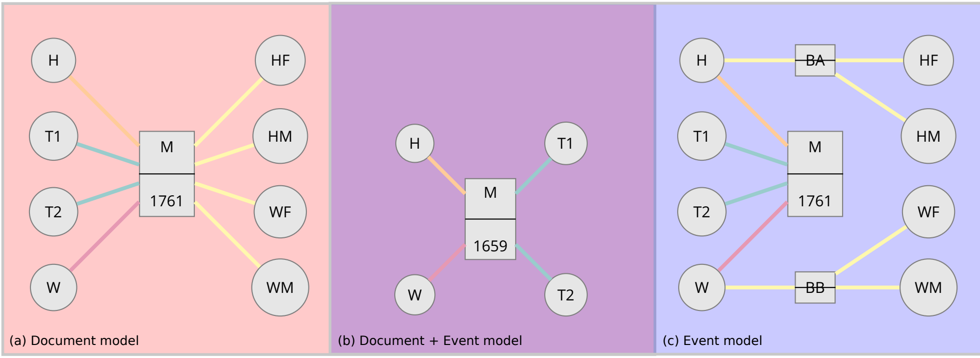


Figure 3.3 – bipartite multivariate dynamic network modeling for two cases of marriage acts of example #3. Some marriage acts mention the parents of the spouses, which is a relationship different than the marriage in itself. This case can be modeled using a document model (a) or an event model (c) by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

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 the resulting networks. We tried to shed light on this process by dividing it into 5 steps and describing recurrent pitfalls we encountered in our projects and collaborations. More importantly, we think this process should be done following the principles of *reality*, *traceability*, and *simplicity*, to avoid biasing the analysis, allowing to go back to the original source at any point of the workflow, and using models and methods simple and powerful enough for social scientists. Visual analytics software designed for HSNA should consider those principles to provide tools allowing to follow non-biased and reproducible analysis starting from the raw documents while supporting historians in going back and forth more easily between the annotation and analysis/visualization steps. We discussed the network modeling process in depth and claim that bipartite multivariate dynamic networks satisfies those three core principles, letting historians both wrangle their data and characterize sociological phenomena using a common model and visual representation. Therefore, using this model VA interfaces could help social scientists manage and analyze their data starting at the data acquisition and annotations steps instead on focusing on the analysis only, while providing efficient representations of the data for analysis and exploration. We explore what could be such VA interfaces in the two next chapters.

Bibliography

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

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

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

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

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

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

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

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

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

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

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

[12] Christian Bors, John Wenskovitch, Michelle Dowling, Simon Attfield, Leilani Battle, Alex Endert, Olga Kulyk, and Robert S. Laramee. A Provenance

Task Abstraction Framework. *IEEE Computer Graphics and Applications*, 39(6):46–60, November 2019.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[36] Cody Dunne, Nathalie Henry Riche, Bongshin Lee, Ronald Metoyer, and George Robertson. GraphTrail: Analyzing large multivariate, heterogeneous

networks while supporting exploration history. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1663–1672, New York, NY, USA, May 2012. Association for Computing Machinery.

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

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

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

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

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

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

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

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

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

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

[47] GEDCOM: The genealogy data standard.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[67] Infovis SC policies FAQ.

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

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

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

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

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

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

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

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

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

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

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

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

[80] Claire Lemercier. 12. Formal network methods in history: Why and how? In Georg Fertig, editor, *Social Networks, Political Institutions, and Rural Societies*, volume 11, pages 281–310. Brepols Publishers, Turnhout, January 2015.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[99] Natural earth.

[100] Neo4j graph data platform.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[123] Anni Sairio. Methodological and practical aspects of historical network analysis: A case study of the Bluestocking letters. In Arja Nurmi, Minna Nevala,

and Minna Palander-Collin, editors, *Pragmatics & Beyond New Series*, volume 183, pages 107–135. John Benjamins Publishing Company, Amsterdam, 2009.

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

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

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

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

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

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

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

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

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

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

[134] SNA — Tools for social network analysis.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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