From Historical Documents To Social Network Visualization: Potential Pitfalls and Network Modeling

Alexis Pister, Nicole Dufournaud, Pascal Cristofoli, Christophe Prieur, Jean-Daniel Fekete

From Historical Documents To Social Network Visualization: Potential Pitfalls and Network Modeling

Alexis Pister*
Université Paris-Saclay, CNRS, Inria
Institut Polytechnique de Paris

Nicole Dufournaud†
Université Gustave Eiffel

Pascal Cristofoli‡
EHESS

Christophe Prieur§
Université Gustave Eiffel
Institut Polytechnique de Paris

Jean-Daniel Fekete¶
Université Paris-Saclay, CNRS, Inria

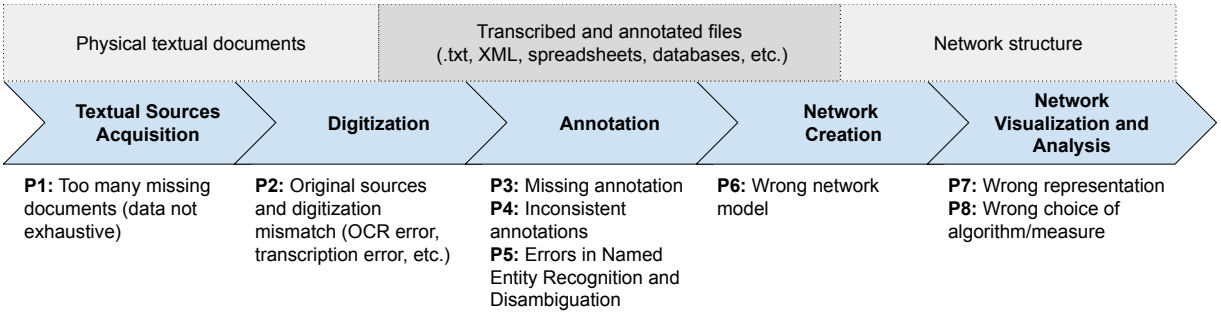

Figure 1: HSNA workflow split into five steps: textual sources acquisition, digitization, annotation, network creation, network visualization, and analysis. We list potential pitfalls for each step.

ABSTRACT

We describe the workflow followed by historians when conducting a Historical Social Network Analysis (HSNA) with five steps: textual sources acquisition, digitization, annotation, network creation, and analysis/visualization. While most analysis and visualization tools only support the last step, we argue that addressing the 2–3 last steps would boost the humanists' analytical capabilities. We explain why the network modeling process is particularly challenging and can lead to distortions of the sources, biases, and traceability problems. We list three main properties that we believe the constructed network should satisfy: alignment with reality/documents (not only with concepts), traceability (from documents to analysis/visualization and back), and simplicity (understandable by most and not more complex than needed). We claim that the model of *bipartite dynamic multivariate network with roles* allows an effective annotation/encoding of historical sources while satisfying these properties. We provide real-world examples of how this model has been used to answer socio-historical questions using visual analytics tools.

Index Terms: Applied computing—Arts and humanities Human-centered computing—Visualization—Visualization application domains—Visual analytics;

1 INTRODUCTION

Tools for social network visualization tend to ignore the context in which the networks are produced, where they come from, and the workflow that led from their origin (e.g., documents, polls, interviews, web scraping) to their network form. Yet, social historians need to generate many networks from the same documents/sources to visualize and analyze them. In this article, after describing the current Historical Social Network Analysis [48] (HSNA) workflow, we explain why and how effective tools for supporting this process should model social networks in multiple steps to support three essential principles: *traceability*, connection to *reality*, and *simplicity*. These principles emerged from our joint experience 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, to see how individual people of that time lived through those changes. For this, they rely on historical documents such as conversational letters, censuses, and marriage acts. They usually extract qualitative and quantitative information from an identified corpus of documents, to then make conclusions on interesting socio-economic topics such as migrations, business dynamics, education, and kinship. For doing this, historians can apply Social Network Analysis (SNA), a method—sometimes referred to as a paradigm—which consists in modeling the social relationships between a set of entities—usually individuals—into a network. Much work has been done to adapt SNA to the context of historical document exploitation, and although several approaches coexist, they can be brought together under the banner of Historical Social Network Analysis [48] (HSNA) or Historical Network Research [27] (HNR). While the goals of HSNA and SNA are similar, several important differences exist. Sociologists can construct a network from direct observations of the world; historians only rely on sources, often incomplete. They collect dated documents, annotate them, and construct networks from the annotations that they finally analyze and visualize to find hypotheses or invalidate them. Their network model is strongly influenced by the structure of the documents. Moreover, historians always need to relate to the temporal aspect of their sources that should be reflected in the networks, which is not always the case in sociology.

Unfortunately, the HSNA 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.

*e-mail: alexis.pister@inria.fr
†e-mail: nicole.dufournaud@univ-eiffel.fr
‡e-mail: pascal.cristofoli@ehess.fr
§e-mail: christophe.prieur@univ-eiffel.fr
¶e-mail: Jean-Daniel.Fekete@inria.fr

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

2 Related Work

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

2.1 Quantitative History

Traditionally, historians try to tell a story about protagonists and socio-economic facts in a given society by reading, understanding, and linking together historical sources. This narrative approach to history has been criticized for its lack of traceability and the open interpretation of historical documents, which can introduce bias from the author. To solve this problem, the "Annales school" (Ecole des Annales) proposed to characterize past social phenomena through the exhaustive and systematic analysis of historical documents [41]. Quantitative approaches then emerged in the 1960s with the appropriation of statistical and computer science methods to analyze data extracted from historical documents. This is the case of Historical Demography, which works on nominative data to produce quantitative results on fertility and mortality [22]. Unfortunately, these approaches have been criticized for their simplifications and for consuming considerable time while often providing simple results [26]. Approaches using digital methods and tools are nonetheless more and more popular, sometimes referred to under the umbrella term Digital Humanities. If their adoption remains slow and sometimes criticized among historians, they still provide tools to store, explore, and analyze historical documents systematically if used appropriately (i.e. not trying to bias the analyzes) [8, 28]. It can also provide infrastructure and tools to study large historical databases, as with the Venice Time Machine project [25] which aims at digitizing and analyzing thousands of documents from the archives of Venice to understand the political, geographical, and sociological dynamics of the cities across generations and centuries.

2.2 Social Network Analysis and History

In Sociology, networks have been a common metaphor to talk about social relationships [16] which can be easily thought of as invisible bonds linking individuals, forming a global structure of connections similar to a mesh or a web. After extensive work in graph theory and network modeling developed in the 1950s, anthropologists and sociologists started to borrow those concepts from maths and use them to model social relations—such as family, friendships, or business ties—with networks [7, 16]. SNA revolutionized classical Sociology by trying to explain social phenomena through the lens of real interactions modeled as networks [1], while classical methods were revolving around predefined social groups such as age and gender.

History started to use those concepts and methods in the 1980s [48] in order to criticize quantitative history concepts and results, and to develop historical approaches—like *Microstoria* [18]—that focus on the study of individuals and groups through the lens of their interactions and relationships directly extracted from historical documents. Since then, HSNA has been applied by sociologists and historians to study multiple kinds of relationships, like kinship and political mobilization [31], administrative and economic patronage [34], etc. If these approaches fall under similar critics of quantitative history [29], lots of historians are using and continuously improving this method which can be very effective in studying relational historical phenomena [27]. Moreover, historians rarely rely on a single approach when studying an era or phenomenon, they mix methods and tools from several domains of social and formal sciences with their own practices [37, 39].

2.3 Network Modeling

Social scientists started to model social relationships using simple graphs $G = (V, E)$ with V a set of vertices representing actors—very often persons—and $E \subseteq V^2$ a set of edges modeling a social tie between pairs of actors [16].

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

Historians, demographers, sociologists, and anthropologists have been designing specific data models for their social networks, based on genealogy or more generally, kinship [21]. For genealogy, the standard GEDCOM [17] format models a genealogical graph as a bipartite graph with two types of vertices: individuals and families. This format also integrates an "event" object, but it is diversely adapted in genealogical tools. The Puck software has extended its original genealogical graph with the concept of "relational nodes" to adapt the data model to more family structures and to integrate other social relationships for anthropology and historical studies [20].

2.4 Social Network Visualization

Social scientists such as sociologists and historians always used visual representations for social networks [10], mainly for communication purposes and sometimes for exploration [6]. Moreno elaborated on sociograms in the 1930s to visualize friendships using circles and lines to represent persons and ties, in a node-link fashion [33]. Node-link diagrams are still the most widely used technique in SNA and HSNA by far to represent networks, despite scalability issues. The most used social network visual analytics software, such as Gephi [4] and Pajek [35], are based on this type of representation and allow a fully integrated exploration and analysis with the help of various algorithms. The visualization community also proposed other representations to visualize networks, such as matrices [5], and to explore other network types such as dynamic hypergraphs with PAOHVis [46], clustered graphs with NodeTrix [23], geolocated social networks with the Vistorian [43], and multivariate networks with Juniper [36]. Jigsaw [44] is designed to analyze a

collection of documents. It encompasses the gathering of documents, named entity recognition, with analysis and visualization methods; the last three steps of our workflow. These documents and entities form a multi-partite network with a few attributes (e.g., document title, date) but without roles. Except for Jigsaw, most of the tools proposed are solely focused on the exploration and analysis of the final network and do not take into account the context of the whole HSNA process which led to the network creation.

3 Historical Social Network Analysis Workflow

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 original sources. Social history and the "Annales School" brought answers to this problem by proposing to rigorously extract information from historical documents and make conclusions from them. Similarly, Glaser and Strauss developed the "Grounded Theory" [19] as a methodology for the humanities to build hypotheses and theories by solely studying and categorizing real-world observations, without starting from prior knowledge and predefined categories. Later on, in the 1960s, quantitative methods started to be used in history, providing statistical and later computer-supported tools to aid historians in grounding their analysis in mathematical models and methods. Unfortunately, the lack of methodology and understanding between the two worlds led to many criticisms by historians pointing to using wrong metrics, simplifying categories, and disconnections between the original documents and analysis [26, 30]. HSNA, which can be seen as a sub-component of quantitative history, has been criticized for similar reasons [29]. Still, the usage of networks for historical analysis provided some interesting results and classical works [37], meaning that a clearer and more rigorous methodology with simple tools grounded in the historian workflow and sources could improve the methods of the field. Karila-Cohen and al. provide advice on how to use quantitative methods in history [26] while Dufournaud describes her workflow when studying the socio-economic status of women in France in the 16th and 17th centuries, which she splits into three steps: *data collection*, *data processing*, and *data analysis* [14]. From 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* (see Fig. 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 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 needed human work to be digitized. **Mismatches between the original documents and the transcription can occur for old and recent documents (P2)**. However, if OCR tools are more and more efficient in English and highly used languages, historians can work with old documents written in old or extinguished languages and with atypical writings (e.g., Fraktur handwriting and typefaces for German in the early 20th century). Therefore, OCR tools are often unusable in social history, and digitization remains an expensive and sometimes highly skilled process.

Annotation Annotation is the process of finding and extracting useful information from the documents concerning the persons, their social ties, and any useful information for the historian. This extra information can concern the persons (their age, profession, sex, ethnicity, etc.) and their social relationships (type, date, place). It encompasses named-entity recognition (NER) as well as their resolution. Historians also sometimes annotate information on other entities mentioned in the documents, such as art objects or administrative entities. Usually, historians have a first idea of what they want to annotate in the data as they already explored the documents beforehand and have knowledge of their subject of study, with hypotheses they want to explore. It is, however common, that they can change their mind through the annotation process, by reflecting on what they found in the documents. Unfortunately, this can produce **missing annotations (P3)** and **inconsistent annotations (P4)** at the end of the process if annotators are not careful. This task can also be challenging, and the choice of annotations has an impact on the final network. Historians also face ambiguity in the process, as several different persons and entities (like cities) can have the same name (homonyms), refer to a place name that has disappeared (street name or city), or to an ambiguous person (e.g., John Doe). They, therefore, have to follow a NER and resolution/disambiguation process to identify entities in the sources and disambiguate them across several documents. Entity resolution has always been a problem in social history—as it is more generally in text analysis, where typical groundwork consists in crossing information about the same entities from different heterogeneous sources [3]. However, errors in the disambiguation process can lead to important distortions in the final network structure and properties [13], 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) [45] 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 [43]. Unfortunately, the guidelines are not meant to define a canonical annotation, and different persons can interpret the guidelines in different ways, leading again to inconsistent annotations of corpora (P4) and to errors or distortions in social networks derived from these annotations.

Network Creation Historians construct a network from the annotations of the documents. Usually, all persons mentioned are annotated and will be transformed into network nodes (vertices). Additional information such as their age, profession, and gender can be stored as node attributes. How the network's links are created is not as trivial and can vary from project to project [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.

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 [16]. Usually, historians start to represent their network to visually confirm information they know, then to gain new insight with exploration. Representations need to be chosen wisely given the network as **some insight may be seen only with some specific visualization technique (P7)**. To test or create a new hypothesis, historians usually rely on algorithms and network measures. **They have to interpret the results carefully (P8)** as some algorithms act as black boxes and some measures are hard to interpret, with unclear sociological meaning (e.g., centrality).

4 Network modeling and analysis

Historians usually construct one or several networks from their annotated documents that they will visualize and analyze to validate or find new hypotheses. As the processing steps of the workflow are often not transparent (digitization, annotation, network modeling), it can be difficult for the reader of an HSNA study to understand how the network has been constructed, what it represents, and to trace back the network entities to the original sources [14]. 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. Moreover, network models satisfying *traceability*, *reality* and *simplicity* properties would mitigate those problems by allowing to navigate more easily between the network and the documents while still modeling well the social relationships mentioned in the sources and being easy enough to visualize and manipulate to detect potential errors and inconsistencies.

4.1 Network Models

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

- **Simple Networks [48]:** 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 [42]:** 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 [33]. However, historical documents rarely mention only one type of relationship, and this model is thereby very limiting for HSNA.
- **Multiplex Unipartite Networks [15] :** Only the persons are represented as nodes, and links model social ties between two persons. Links can have different types representing different types of social relationships e.g., as parent, friends, and business relationships. One of the main drawbacks of this model is that it creates parallel edges that are hard to visualize.
- **Bipartite (also called 2-mode) Networks [20] :** 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 [11]. 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 [32]:** in these networks, each node 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 [12] and historians [47], but they are complex. The meaning of a layer varies from one application to another; it can be time (years), type of document, origin of the source, etc. They, therefore, offer many options for modeling a corpus, and visualizing it, with no generic system to support historians in taming their high complexity.
- **Knowledge Graphs (KG) [24]:** 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 little support for taming their high complexity.

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.

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 [44] 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.

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

One of the main benefits of this model is that the document nodes represent both the physical documents as well as the events the documents refer to. For example, concerning marriage acts, the document nodes represent both the physical documents with their texts and also the marriage events with their characteristics modeled as attributes (time, location, etc.). This model is *simple* enough to manipulate and visualize for historians and allows tracing back every entity of the network to the documents according to the *traceability* principle. Still, the network preserves the *reality* of the social relationships mentioned in the sources. More precisely, Cristofoli demonstrates that bipartite networks ensure no distortion or ambiguity, unlike projected networks when modeling textual sources [9]. Furthermore, when attributes are encoded, projected networks require the duplication of information related to events. For example, the date of a marriage can be encoded in the document/event node with a bipartite network while the same information would have to be stored in parallel links when using a projection.

5 Applications

Several tools have been designed for visualizing dynamic bipartite networks that can also be considered dynamic hypergraphs [38, 46], but few support attributes and complex interactions. We designed ComBiNet [40] to explore and navigate through historical documents modeled as bipartite multivariate dynamic networks and to help social scientists answer their questions with the help of visual queries and interactive comparisons of query results. Fig. 2 shows the interface to compare two meaningful groups of construction documents in Piemont during the 18th-century [11]. In this example, we see that the *Zo* family has more construction contracts in *Turin* than the *Menafoglio* family. Exploring historic datasets modeled as bipartite multivariate dynamic networks allows answering complex questions both related to the events (here the constructions) and the persons while being able to trace back to the original documents directly in the interface for cleaning or debugging purposes. The interface, therefore, facilitates the loop between the visualization and annotation steps, towards new HSNA Visual Analytics environments supporting historians in their digital workflow.

6 Discussion

Most tools for social network visualization focus solely on the visualization and analysis steps, without considering the whole historical analysis process, preventing researchers from going back to the original source, the annotation, or 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 for conducting socio-historical inquiries with limited friction, realistic training, and scientific transparency. Concerning the

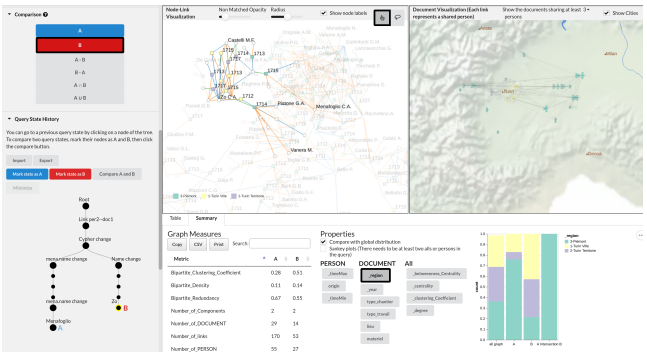

Figure 2: ComBiNet interface exploring construction contracts in Piemont during the 18th century [11]. The left menu lets users filter the data and compare groups. The center view shows the bipartite network with a node-link diagram. The right view shows a map with the geolocated construction contracts. The bottom view gives measures and attribute distributions related to the network and the current filters and comparisons. The user currently compares the *Menafoglio* and *Zo* families in terms of their construction types and close relationships.

network modeling step, the 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 events. For example, marriage acts sometimes refer to the place and date of birth of the spouses along with the names of the parents. In that case, social historians can either ignore this type of information in the annotation process or encode it with specific roles (*husband's father* and *wife's father*, for example), complexifying the network's model with redundant information.

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.

Acknowledgments

This research was supported by DATAIA as part of the "Programme d'Investissement d'Avenir", `ANR-17-CONV-0003` operated by Inria.

REFERENCES

[1] R. Ahnert, S. E. Ahnert, C. N. Coleman, and S. B. Weingart. The Network Turn: Changing Perspectives in the Humanities. *Elements in Publishing and Book Culture*, Dec. 2020.

[2] M. AlKadi, V. Serrano, J. Scott-Brown, C. Plaisant, J.-D. Fekete, U. Hinrichs, and B. Bach. Understanding Barriers to Network Exploration with Visualization: A Report from the Trenches. *IEEE Trans. Vis. Comput. Graphics*, 27(2), Feb. 2023. to appear.

[3] O. Andrei, M. Fernández, H. Kirchner, G. Melançon, O. Namet, and B. Pinaud. Porgy: Strategy-driven interactive transformation of graphs. *arXiv preprint arXiv:1102.2654*, 2011.

[4] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An Open Source Software for Exploring and Manipulating Networks. In E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, and B. L. Tseng, eds., *ICWSM' 2009*. The AAAI Press, 2009.

[5] M. Behrisch, B. Bach, N. Henry Riche, T. Schreck, and J.-D. Fekete. Matrix Reordering Methods for Table and Network Visualization. *Computer Graphics Forum*, 35(3):693–716, 2016.

[6] U. Brandes, J. Raab, and D. Wagner. Exploratory Network Visualization : Simultaneous Display of Actor Status and Connections. *Journal of Social Structure*, 2(4):28, 2001.

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

[8] R. Cordell and D. Smith. Viral texts: Mapping networks of reprinting in 19th-Century newspapers and magazines, 2017.

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

[10] P. Cristofoli. Principes et usages des dessins de réseaux en SHS. *Histoire et Informatique*, 18/19:23–58, 2015.

[11] P. Cristofoli and N. Rolla. Temporalités à l'œuvre dans les chantiers du bâtiment. *Temporalités*, 1(27), 2018.

[12] T. Crnovrsanin, C. W. Muelder, R. Faris, D. Felmlee, and K.-L. Ma. Visualization techniques for categorical analysis of social networks with multiple edge sets. *Social Networks*, 37:56–64, 2014.

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

[14] N. 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 B. Michon and N. Dufournaud, eds., *Femmes et Négoce Dans Les Ports Européens (Fin Du Moyen Age - XIXe Siècle)*, pp. 65–84. Peter Lang, 2018.

[15] E. Erikson and P. 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.

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

[17] Gedcom: The genealogy data standard. `https://www.gedcom.org/`. Last access: Nov. 18, 2021.

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

[19] B. G. Glaser and A. L. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Transaction, New Brunswick, 5. paperback print ed., 2010.

[20] K. Hamberger, C. Grange, M. Houseman, and C. Momon. Scanning for patterns of relationship: analyzing kinship and marriage networks with Puck 2.0. *The History of the Family*, 19(4):564–596, Oct. 2014.

[21] K. Hamberger, M. Houseman, and R. White, Douglas. Kinship Network Analysis. In J. S. . P. J. Carrington, ed., *The Sage Handbook of Social Network Analysis*, pp. 533–549. Sage Publications, 2011.

[22] L. Henry and M. 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.

[23] N. Henry, J.-D. Fekete, and M. J. McGuffin. NodeTrix: a hybrid visualization of social networks. *IEEE Trans. Vis. Comput. Graphics*, 13(6):1302–1309, 2007.

[24] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, and S. Kirrane et al. Knowledge graphs. *ACM Comput. Surv.*, 54(4), jul 2021.

[25] F. Kaplan. The Venice Time Machine. In *ACM Symposium on Document Engineering*, p. 73. ACM, Sept. 2015.

[26] K. Karila-Cohen, C. Lemercier, I. Rosé, and C. Zalc. Nouvelles cuisines de l'histoire quantitative. *Annales. Histoire, Sciences Sociales*, 73(4):773–783, Dec. 2018.

[27] F. Kerschbaumer, L. von Keyserlingk-Rehbein, M. Stark, and M. Düring. *The Power of Networks. Prospects of Historical Network Research*. Routledge, Dec. 2021.

[28] L. Klein and J. Eisenstein. Reading Thomas Jefferson with TopicViz: Towards a Thematic Method for Exploring Large Cultural Archives. *Scholarly and Research Communication*, 4(3), Dec. 2013.

[29] C. Lemercier. 12. Formal network methods in history: Why and how? In G. Fertig, ed., *Social Networks, Political Institutions, and Rural Societies*, vol. 11, pp. 281–310. Brepols Publishers, Jan. 2015.

[30] C. Lemercier and C. Zalc. Back to the Sources: Practicing and Teaching Quantitative History in the 2020s. *Capitalism: A Journal of History and Economics*, 2(2):473–508, 2021.

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

[32] F. McGee, B. Renoust, D. Archambault, M. Ghoniem, A. Kerren, and B. Pinaud et al. *Visual Analysis of Multilayer Networks*. Synthesis Lectures on Visualization. Morgan & Claypool Publishers, 2021.

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

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

[35] A. Mrvar and V. Batagelj. Analysis and visualization of large networks with program package Pajek. *Complex Adaptive Systems Modeling*, 4(1), Apr. 2016.

[36] C. Nobre, M. Streit, and A. Lex. Juniper: A Tree+Table Approach to Multivariate Graph Visualization. *IEEE Trans. Vis. Comput. Graphics*, 25(1):544–554, Jan. 2019.

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

[38] V. Peña-Araya, T. Xue, E. Pietriga, L. Amsaleg, and A. Bezerianos. HyperStorylines: Interactively untangling dynamic hypergraphs. *Information Visualization*, 21(1):38–62, Jan. 2022.

[39] C. S. M. Petz. *On Combining Network Research and Computational Methods on Historical Research Questions and its Implications for the Digital Humanities*. PhD thesis, TU München, 2022.

[40] A. Pister, C. Prieur, and J.-D. Fekete. Visual queries on bipartite multivariate dynamic social networks. In *EuroVis 2022-Posters*, 2022.

[41] A. Prost. *Douze Leçons sur l'histoire*. Média Diffusion, Apr. 2014.

[42] A. Sairio. Methodological and practical aspects of historical network analysis: A case study of the Bluestocking letters. In *Pragmatics & Beyond New Series*, vol. 183, pp. 107–135. John Benjamins, Amsterdam, 2009.

[43] V. Serrano Molinero, B. Bach, C. Plaisant, N. Dufournaud, and J.-D. Fekete. Understanding the Use of The Vistorian: Complementing Logs with Context Mini-Questionnaires. In *Visualization for the Digital Humanities Workshop*. Phoenix, United States, Oct. 2017.

[44] J. T. Stasko, C. Görg, and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Inf. Vis.*, 7(2):118–132, 2008.

[45] TEI Consortium. TEI P5: Guidelines for Electronic Text Encoding and Interchange, Feb. 2021.

[46] P. Valdivia, P. Buono, C. Plaisant, N. Dufournaud, and J.-D. Fekete. Analyzing Dynamic Hypergraphs with Parallel Aggregated Ordered Hypergraph Visualization. *IEEE Trans. Vis. Comput. Graphics*, 27(1):1–13, Jan. 2021.

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

[48] C. Wetherell. Historical Social Network Analysis. *International Review of Social History*, 43(S6):125–144, Dec. 1998.