université PARIS-SACLAY

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

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

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

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

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

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

Alexis PISTER

Composition du jury

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

THESE DE DOCTORAT
NNT : 2020UPASA001

université PARIS-SACLAY

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

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

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

Résumé: Cette thèse vise à identifier comment l'analyse visuelle peut aider les historiens dans leur processus d'analyse de réseaux sociaux, de la collecte des sources jusqu'à la formulation de conclusions socio-historiques. L'analyse de réseaux sociaux historiques est une méthode permettant d'étudier les relations sociales au sein de groupes d'acteurs (familles, institutions, entreprises, etc.) pour comprendre leurs structures sous-jacentes tout en décrivant des comportements spécifiques. Les chercheurs en histoire sociale reconstruisent les relations du passé à partir du contenu de documents historiques, tel que des actes de mariage, formulaires de migration, ou des recensements. Via des méthodes analytiques et de visualisation, les historiens peuvent décrire la structure de ces groupes et expliquer des comportements individuels à partir de motifs locaux. Cependant, l'inspection, l'encodage et la modélisation des sources pour obtenir un réseau finalisé donne souvent lieu à des erreurs, distorsions et des problèmes de traçabilité. Pour ces raisons, ainsi que des problèmes d'utilisabilité, les historiens ne sont pas toujours en mesure de faire des conclusions approfondies sur leur réseau à partir des systèmes de visualisation actuels. Je vise dans cette thèse à identifier comment l'analyse visuelle (la combinaison d'algorithmes statistiques intégrés à des interfaces graphiques à l'aide d'interaction) peut aider les historiens dans leur processus, de la collecte des données jusqu'à l'analyse finale. Dans ce but, je formalise le processus d'une analyse de réseau historique en partant de collaborations avec des historiens, de l'acquisition des sources jusqu'à l'analyse visuelle, en établissant que les outils aidant ce processus devraient satisfaire des principes de traçabilité, simplicité et de réalité documentaire pour faciliter les va-et-vient entre les différentes étapes, avoir des outils faciles à utiliser, et ne pas distordre le contenu des sources. Pour satisfaire ces propriétés, je propose de modéliser les sources historiques en réseaux sociaux bipartis multivariés dynamiques avec rôles. Ce modèle a la particuliarité d'integré une representation des documents historiques, ce qui permet à la fois aux utilisateurs d'encoder, corriger et analyser leurs données avec les mêmes outils. Je propose deux interfaces d'analyse visuelle pour manipuler, explorer et analyser ce type de données, en respectant les principes de traçabilité, simplicité et réalité documentaire. Je présente d'abord ComBiNet, qui permet une exploration visuelle de la topologie, la dynamique, la localisation et des attributs du réseau à l'aide de vues coordonnées et d'un système de requêtes visuelles et de comparaisons.

Le second système, PK-Clustering, constitue une proposition concrète pour améliorer l'utilisabilité et l'efficacité des mécanismes de clustering dans les systèmes de visualisation de réseaux sociaux. L'interface permet de créer des regroupements pertinents à partir des connaissance a priori, du consensus algorithmique et de l'exploration du réseau dans un cadre d'initiative mixte. Les deux systèmes ont été conçu à partir des besoins et de retours continus d'historiens, et visent à augmenter la traçabilité, la simplicité, et la vision réelle des sources dans l'analyse de réseaux historiques. Je conclus sur la possibilité de fusionner les deux systèmes et plus globalement sur la necessité d'une meilleur intégration des systèmes d'analyse visuelle dans le processus de recherche des historiens. Cette intégration nécessite la prise en compte des trois propriétés visées auparavant dans les systèmes, afin de mieux coller aux besoins et aux méthodes des historiens, et ainsi limiter les barrières d'utilisation des méthodes quantitatives, qui subsitent en histoire.

Title: Visual Analytics for Historical Social Networks: Traceability, Exploration, and Analysis
Keywords: visual analytics, social network analysis, social network visualization, social history, historical networks

Abstract: This thesis aims at identifying how Visual Analytics can support historians in their social network analysis process, from the collection of historical documents to the formulation of high-level socio-historical conclusions. Historical Social Network Analysis is a method to study social relationships between groups of actors (families, institutions, companies, etc.) to understand their underlying structure while characterizing specific behaviors. Social historians are able to reconstruct relationships of the past using historical documents' content, such as marriage acts, migration forms, birth certificates, and censuses. Through visualization and analytical methods, they can describe the global structure of studied groups and explain individual behaviors through local network patterns. However, the inspection, encoding, correction, and modeling process of the historical documents leading to a finalized network is intricate and often results in inconsistencies, errors, distortions, simplifications, and traceability issues. For these reasons, social historians are not always able to make thorough historical conclusions with current analytical and visualization tools. I aim in this thesis to identify how visual analytics—the integration of data mining capabilities into visual interfaces with interaction—can support social historians in their process, from the collection of their data to the answer to high-level historical questions. Towards this goal, I formalize the workflow of historical network analysis in collaboration with social historians, from the acquisition of their sources to their final visual analysis, and point out that visual analytics tools supporting this process should satisfy traceability, simplicity, and document reality principles to ease bask and forth between the different steps, provide tools easy to manipulate, and not distort the content of sources with modifications and simplifications. Particularly, I propose to model historical sources into bipartite multivariate dynamic social networks with roles to satisfy those properties. This modeling allows a concrete representation of historical documents, hence letting users encode, correct, and analyze their data with the same abstraction and tools. Leveraging this data model, I propose two interactive visual interfaces to manipulate, explore, and analyze this type of data with a focus on usability for social historians. First, I present ComBiNet, which allows an interactive exploration leveraging the structure, time, localization, and attributes of the data model with the help of coordinated views, a visual query system, and comparison mechanisms. Finding specific patterns easily and comparing them, social historians are able to find inconsistencies in their annotations and answer their high-level questions. The second system, PK-Clustering, is a concrete proposition to increase the usability and effectiveness of clustering mechanisms in social network visual analytics systems. It consists in a mixed-initiative clustering interface that let social scientists create meaningful clusters with the help of their prior knowledge, algorithmic consensus, and interactive exploration of the network. Both systems have been designed with continuous feedback from social historians, and aim to increase the traceability, simplicity, and document reality of visual analytics supported historical social network research. I conclude with discussions on the potential merging of both systems and more globally on research directions towards better integration of visual analytics systems on the whole workflow of social historians. Such systems with a focus on those properties—traceability, simplicity, and document reality—can limit the introduction of bias while lowering the requirements for the use of quantitative methods for historians and social scientists which has always been a controversial discussion among practitioners.

Aknowledgments

On the usage of the pronouns we and I

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

Contents

List of Figures

List of Tables

1 Introduction

> *"My claim rests on the assumption that [...] researchers can learn the truth about social processes. At a minimum, they can distinguish between totally inadequate and less inadequate representations of social processes, thus opening the way to increasingly reliable knowledge."*
>
> **-Charles Tilly, [191]**

The goal of this thesis is to characterize and produce visual analytics tools that can support social historians conducting research on their sources—particularly when using network methods—with a focus on exploration, analysis, traceability, and usability. Historical Social Network Analysis (HSNA) is a method—sometimes referred as a paradigm [205]—followed by social historians to study sociological phenomena through the observation of relationships of actors of the past, modeled into a network. The usage of networks as an abstraction to represent and study social relationships—such as friendships, kinship, or business ties—grew in popularity in the last 40 years [71, 187] and constitute a powerful metaphor, especially in our time when many of our digital connections and interactions use an explicit network structure[1]. This approach has first been formalized in sociology under the term Social Network Analysis (SNA) [71] and is now widely used in anthropology [**?**], geography [**?**], and history [104]. Historians leverage historical documents—which are at the core of their profession [111]—to extract relationships between actors of interest that they model with networks constructed from nodes and links that respectively represent actors (often persons) and relationships (like kinship). Using social network visualization techniques and leveraging network measures and computations, they can then test hypotheses they have and gain insight on the structural aspect of the relational phenomena they are studying [104, 203]. This approach has been followed successfully to study various subjects such as kinship [87], entrepreneurship [167], maritime routes [112], political power [145], political oppositions [144], and persecution [127]. Yet, history is considered by many as a literary and qualitative science, and many critics emerged from the history community concerning quantitative and network methods [85, 101, 116, 119], pointing to problems such as the leading to trivial conclusions, anachronisms, simplifications, and mismatches between network and historical concepts. Moreover, quantitative and network analysis are complex processes, and demand many efforts in data collection, encoding, modification, and processing before being able to make efficient observations. This thesis considers the whole workflow of social historians to better support it with visual analytics.

Social historians have to take many annotation (sometimes called encoding) and modeling decisions, concerning *what* to model from their sources into a network, and *how* to model it [42,54], i.e., should the information of interest be represented as a node, a link, an attribute, or not reflected in the network at all, and what format should be used. Practically, they typically use ad hoc processing and analysis scripts to transform historical documents to analyzable networks,

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

which is time-consuming, sometimes to end up with trivial or hard to interpret results [5]. Still, HSNA led to many highly regarded studies with thorough conclusions, such as the study of families of power in Florence by Padgett and Ansell where they explained the rise of the Medici family through its central position in the economical, political, and trading networks of powerful families [145] or Gribaudi and Blum work on the social and professional shift during the 19th century in France [84].

The usage of visualization to graphically display networks is common in SNA[2] as it allows to unfold the structure of networks to the eyes, thus letting social scientists confirm hypotheses they had when collecting and exploring their data as well as gaining new insight through the discovery of interesting patterns and trends [43]. Images of networks also constitute an efficient mean of communication, especially in scientific productions [70]. Many visualization techniques and softwares have thus been developed since the birth of SNA, but most popular tools are usually not designed for historians specifically, meaning that they do not regard on the provenance and process leading to the network, and focus on analysis aspects only. Moreover, they usually enforce simple network models without proposing exploration mechanisms, beyond allowing to look at the network structure and computed measures. In result, many HSNA studies show a plot of their network and describe it qualitatively, often by identifying the central actors—sometimes with the help of centrality—but do not go beyond that [117]. *In this thesis, I therefore investigate how visualization can support social historians in their work, first during the pre-analysis process and secondly during the analysis step, with the right level of expressiveness, usability, and traceability.*

1.1 Social History and Historical Social Network Analysis

Social History has continuously evolved since its beginning in the 1930s, especially with the rise of quantitative and network methods based on the development of computer science during the end of the 20th century. If these computer-supported methods are now widely used in history [104, 150], they attracted many criticism from the start—some are which still relevant.

We can trace back the birth of Social History with the formation of the "Annales School" in the 1930s, where historians gained interest in socio-economic questions and started to rely heavily on the exhaustive extraction and analysis of historical documents coming from archives [20, 156]. Beforehand, History was mainly political and event-centered, as the majority of work consisted in narrating and characterizing specific events—such as wars and diplomatic alliances—while eliciting their causes and consequences, and describing the lives of historic figures, such as sovereigns [156]. Social History shifted the focus by aiming to link together sociological, economical, and political issues and by placing individuals at the center of these questions [190]. Later on in the 1960s, with the development of computer science, historians started to use quantitative methods to analyze data extracted from historical documents and

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

make conclusions grounded in statistical results, in various subjects such as demographics [93] and economics [81]. Around the same time, the use and study of networks started to become popular in various disciplines to study real-world relational phenomena based on mathematical computations and measures, especially in sociology and anthropology [32]. A network is an abstraction based on graph theory concepts which can be used to model phenomena based on relationships (called links) between entities (called nodes).

Sociologists appropriated this concept to model social relationships between agents of interest, allowing them to study the sociological structure of groups of interest—such as families, institutions, and companies—and concepts like friendship, oppression, and diffusion using real world observation and mathematical computations. This SNA approach allows analysts to ground results in formal network measures and metrics based on real observations instead of relying on traditional social categories such as age, job, and gender [71]. This shift in the object of study from traditional social classes and aggregates to the observation of relationships of individuals remind the microhistory movement [77] which theorized that following the life of single individuals and small groups enable the making of higher level conclusions about the social structures they live in. Social historians followed this tradition and started to appropriate network concepts to study relational aspects of the past and formalized it under the term Historical Network Research or Historical Social Network Analysis [203]. However, historians do not have the possibility to run surveys or directly observe interactions of the past and are thus constrained by the information contained in historical documents they find in archives. These documents can be anything mentioning social relationships between actors of interest, such as marriage acts, birth certificates, census, migration acts, business transactions, journals. After selecting a corpus of documents, they typically read and inspect in depth several documents while taking notes to have a deeper insight on the content of the sources, which allow them to start eliciting hypotheses. Following this exploration phase, they manually annotate each document and encode the desired information—the mention of persons and their social relationship in the case of a network analysis. This is a long and tedious process that can result in small to large networks that they analyze using network measures to make conclusions on the structure of social groups or social behaviour of individual of interests. Figure 1.1 shows for example an original business document of the 17th century from Nantes (France). The historian have to inspect these documents in depth, extract useful information, and cross-reference the sources to do her quantitative analysis afterwards. The investigation and reading of the historical documents is therefore an exploratory process, where historians start to generate sociological hypotheses from the continuous extraction of insight and revelations of this process, similarly to grounded theory [79]. Once they finalised a network, they can test their hypotheses using qualitative or quantitative methods—based on statistical and network measures. Lemercier and Zalc write "Although history is not an exact science, counting, comparing, classifying, and modeling are nevertheless useful methods for measuring our degree of doubt or certainty, making our hypotheses explicit, and evaluating the influence of a phenomenon." [117] Social historians, therefore, have hypotheses about their subject of study, that they can back up or refute with the help of quantitative and network results, in a way similar to the competing hypotheses workflow of Intelligence Analysis [51]. By pointing to evidence supporting or refuting hypotheses, they

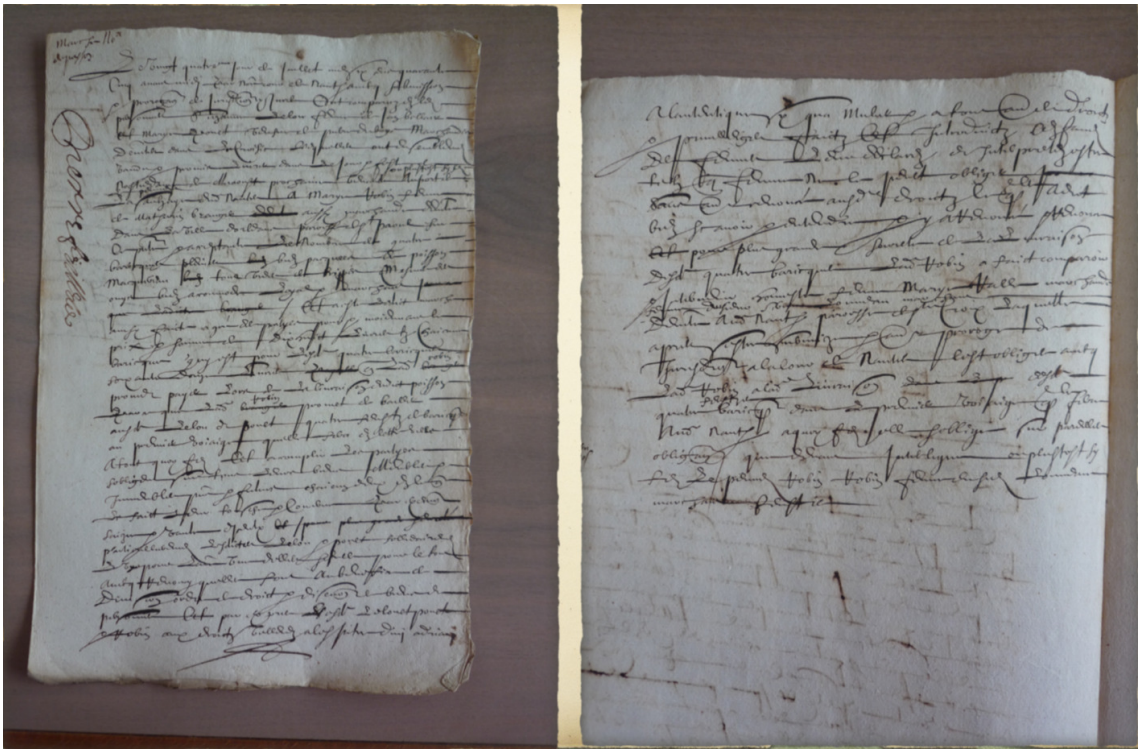

Figure 1.1 – Business contract originated from Nantes (France) during the 17th century. See [55] for more detail of the historian process to analyze her sources.

can give insight into the level of the plausibility of different claims.

1.2 Visualization and Visual Analytics

Visualization has been said to be a central part in the development of SNA [70, 209]—as it the case for many scientific fields[3]. Social scientists now widely use visual and analytical tools to unfold their network structure, allowing them to confirm or deny hypotheses, or follow exploration analysis.

Visualization is the process of displaying data visually to leverage the human visual system and enhance cognition to gain insight into data [35]. Using visual abstractions (such as size, color, and position) to display abstract data allows us to rapidly see structure and patterns otherwise hidden in raw text and numbers. As data keeps growing in size with time due to the increase of hardware and storage capabilities, visualization is a powerful tool to gain insight into the underlying structure of various complex datasets.

Visualization has traditionally been used for confirmatory and communication purposes, particularly in empirical sciences [179]. By showing data visually, analysts are able to confirm or refute hypotheses and communicate their findings in scientific productions.

[3] the historian Alfred Crosby went as far as claiming that visualization is one of the two factors—with measurement—which led to the development of modern science [46].

However, visualization can also be used for exploration, which can help to understand the underlying structure of data and generate new hypotheses. Tukey defined this process as Exploratory Data Analysis in the 1960s [195], as a procedure to gain insight into the structure of the data by identifying outliers, trends, and patterns with the usage of visualization and statistical measures. Social network visualization is used for communication of findings in the field, but is also often following this exploration process as showing the network visually allows social scientists to reveal the structure of their data. As freeman writes "Images of social networks have provided investigators with new insights about network structures and have helped them to communicate those insights to others" [70]. Social scientists very often represent their data using node-link diagrams, that we find in every production of reference in the field [114, 187, 202].

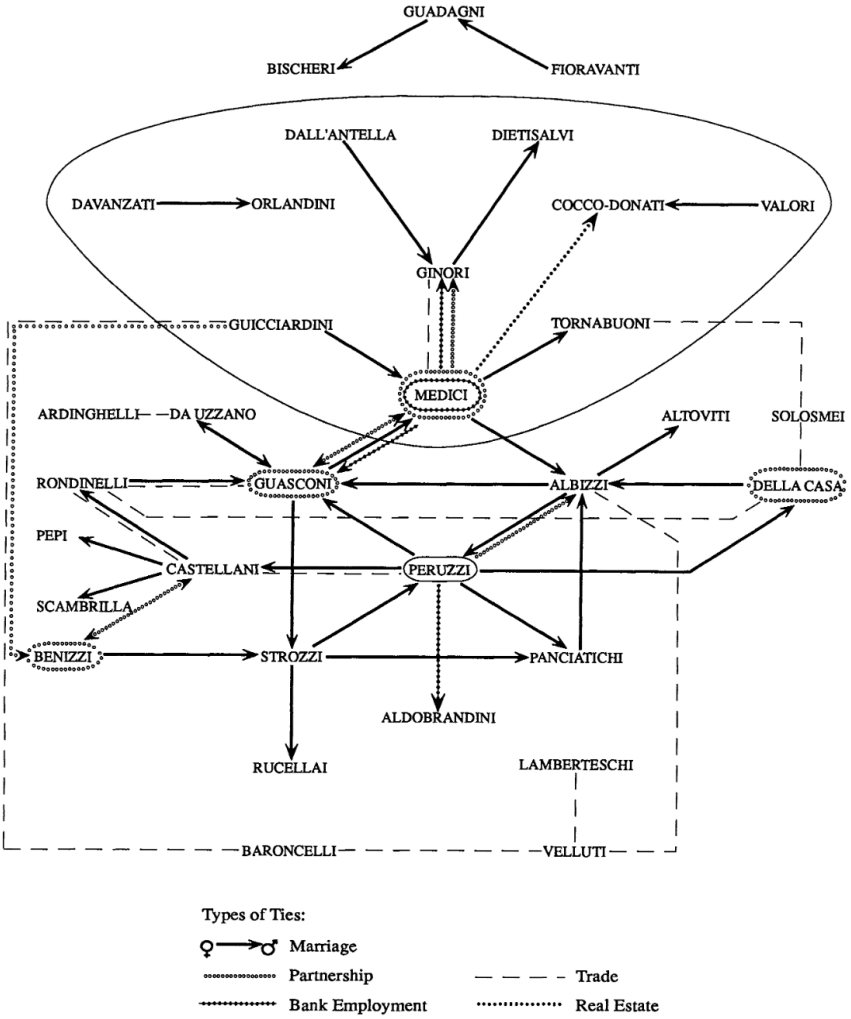

Figure 1.2 – Marriage, partnership, trading, banking, and real estate networks of the powerful families of Florence from [145]. We can see the central position in the network of the Medici Family.

Figure 1.2 shows a node-link representation of the network constructed by Padgett and Ansell in their work on the Medici. At that time, diagrams were often drawn by hand, practice which have now been replaced by automatic layout algorithms. Most visual software for SNA such as Gephi [13], Pajek [146], NodeXl [181], or Ucinet [22] are based on this representation, and allow an exploration of the data with the help of basic interaction mechanisms and the computation of network measures. The detection of patterns and trends can also be facilitated with automatic methods coming from data mining and machine learning fields, directly implemented in the visual analysis loop. This coupling of visual exploration and automatic data mining algorithms has been coined as Visual Analytics (VA) and is defined as the process of using interactive visualizations, transformations, and models of the data in an interactive analysis workflow to create knowledge [102].

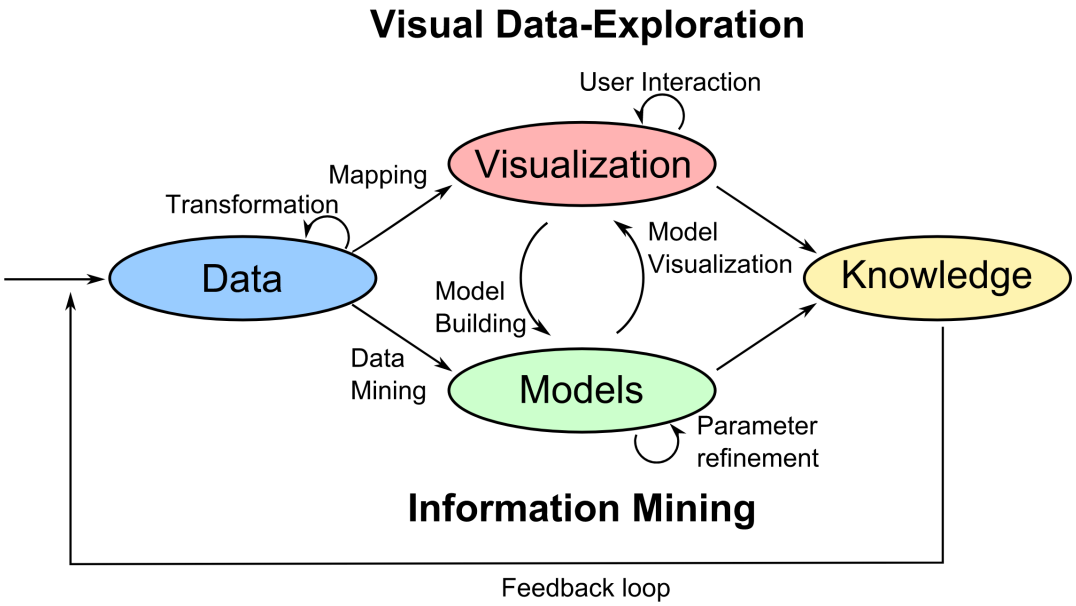

Figure 1.3 – Abstraction of the VA process. It is characterized by continuous interactions between the data, visualizations, models, and knowledge. Image from [102].

Figure 1.3 illustrate the schematic process of VA: the coupling of visualization and data mining models operated by the user through interaction lead to the generation of knowledge "extracted" from the data. If most widely used visual interface for HSNA do not yet provide complex interactions or high data mining capabilities, more recent tools are oriented towards VA, as the combination of automatic knowledge extraction with interaction and exploration can be a powerful support for social scientists to gain insight on the structure of their network, especially that the data they study keep growing in size and complexity [100].

1.3 Visual Analytics Supported Historical Network Research

Most visual tools for SNA are designed for the analysis of already curated networks, without taking into account the context in which those networks have been produced, where they come from, and the workflow that led to their creation. Moreover, many practitioners have trouble using current computer-supported tools, due to misconception in their encoding and modeling process or usability problems [5]. VA should therefore support social historians in the entirety of their process, with a focus on usability and simplicity.

Currently, social historians spend a long time in their data acquisition, processing, encoding, and modeling steps which lead them to the construction of a network [55, 118]. They typically visualize and analyze their network at the end of this process, first to verify hypotheses they formulated during the inspection of their sources, then to gain a better view of the structure of the network, allowing them to potentially generate new hypotheses [116]. However, research showed that all the steps preceding the analysis can introduce errors and misconceptions, especially since social scientists are often not trained in computer science and data science [5, 117]. Social scientists usually visualize their network using SNA tools like Gephi, Pajek, and NodeXl which encompass basic interactions, node-link visualization, SNA measure computations, and clustering algorithms. Once they visualize their data, they typically notice errors and inconsistencies in the data, such as duplication of the same entities, merging of different entities, or geolocation errors [5, 52].

Practitioners also have to decide on a network model [42] (see §2.3.4 for more details) when encoding their documents, which sometimes do not match the final analysis goals. Simple models typically oversimplify the relationships contained in the sources [116] and too complicated models are hard to manipulate [143]. They, therefore, have to go back and forth between the visualization software and the encoding process which can be tedious, especially since it can be complicated to trace back the entities of the data model back to the original documents for correction. VA tools that encompass the whole process of social historians should therefore be beneficial for the flow of their work and could help detect and correct errors or analysis plans way before the visualization of a finalised network. Proposing how to design such interfaces with proof-of-concepts is one of the goal of this thesis.

Furthermore, several historians highlighted the fact that many social history studies leveraging network methods simply use networks in a metaphorical sense, in what Rollinger calls "soft SNA" or "informal network research" [163]. Such studies typically show one—or a couple—node-link diagram(s) which they describe with qualitative terms [117] to refer to the global structure of the network (dense, parse, connected, etc.), the place of actors (central, distant), or interesting patterns (cliques, bridges, communities). In case of dense networks, such descriptions become obsolete, as diagrams start to look like what have been called a "spaghetti monster" [39, 117] i.e., an unreadable image due to the high level of cluttering. Figure 1.4 shows for example a medieval social network of peasants proximity relationships between 1250 and 1350, extracted from agrarian contracts. The graphic do not convey much information, especially that the links represent a constructed notion of proximity without indicating the types of relationships the individuals were mentioned in the contracts.

Figure 1.4 – Node-link diagram of a medieval social network of peasants, produced with a force-directed layout, commonly used in SNA softwares. Image from [26].

The lack of use of network analytical methods—which are numerous in modern SNA softwares[4]—have been in part explained by "math anxiety" [148]: it takes long effort to learn the mathematical concepts behind network measures and algorithms, and their relationships to sociological concepts [163], especially for practitioners without formal computer science and mathematical training. My claim is that current HSNA tools do not support social scientists enough in their analysis due to 1) the lack of interaction, direct manipulation, and exploration mechanisms in current interfaces and 2) the lack of network measures and algorithm interpretations and explainability. For example, clustering algorithms are often included in such

[4]See for example the long technical manuals of Pajek [146] and Ucinet [99]

systems, letting social scientists partition networks into groups, but many algorithms exist in the literature, potentially giving diverse results. Scientists often run several algorithms until finding a satisfying enough partition, which can bias the result of an analysis [153]. Usability and traceability of the results are therefore primordial in VA interfaces aimed at supporting social historians in their analysis.

VA could therefore help social historians using network methods for their research, first by supporting their entire workflow to help them explore, encode, correct, and model their data with simple tools and without introducing oversimplifications, but also to provide guidance and exploration mechanisms during the purely analytical step. For this, such interfaces should therefore 1) be simple enough to manipulate, 2) model the original documents and annotations without distortions, and 3) let historians trace back their network entities to the original sources and analytical results in explainable frameworks. In other terms, they should satisfy *simplicity*, *document reality*, and *traceability* principles. We discuss and explain them more in depth in chapter 3.

1.4 Contributions and Research Statement

The goal of this thesis is to characterize how VA can support social historians in their HSNA process and present proofs of concepts of tools supporting it. Most social network visualization tools are agnostic to the process of social historians leading to a polished network, even though it has an high impact of the network model and structure. Using visualization only at the end of the process often reveals potential errors, inconsistencies, or mismatches between the network model and analysis goals [5]. Moreover, due to lack of usability and interaction mechanisms, social historians often simply visualize statically their network and partially describe their structure, leading to conclusion which would have been easier to reach with simpler methods [61]. VA could therefore 1) assist social historians in their overall workflow, starting at the documents' acquisition to the final analysis step, with the help of data mining and interaction mechanisms in the data acquisition, encoding, modeling steps, and 2) provide exploration and analysis mechanisms to answer complex historical questions, beyond simply plotting the network with a node-link diagram.

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

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

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

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

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

5 PK-Clustering

Contents

As discussed in chapter 1, most SNA tools propose the computation of network measures and data mining algorithms such as clustering. Yet, social scientists are not always in a position to use them efficiently due to interpretability issues, and can become frustrated with automatic results if they do not match their prior knowledge. In this chapter, I address **Q3** by proposing a mixed-initiative approach for network clustering based on the prior knowledge of social scientists, consensus of algorithms, and exploration capabilities. In this framework, social historians are

able to leverage algorithmic power in support of their analysis through interaction while limiting the introduction of bias with reports of actions leading to the final clustering. The system focus on traceability, simplicity, and document reality principles, by respectively reporting the choices leading to the constructed clusters, simple interaction mechanisms, and by leveraging bipartite multivariate dynamic networks as a data model.

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

5.1 Context

In contrast to the belief that most data is easily available on the Web, as of today, most social scientists spend a long time collecting data, to construct social networks, based on documents or surveys, in order to create and carefully validate medium-sized networks (see chapter 3). Before the start of the cluster analysis a great deal of effort goes into analysing other data and gathering knowledge—which I refer under Prior Knowledge (PK) in the rest of the chapter. Social scientists study in great details the network entities (most of the time people), and the social ties they weave together, as it is the unit brick with which they can make historical or social hypothesis and conclusions. When the network is small, less than 30–50 nodes, it is possible to remember most of the relations and persons and visualization directly helps to show groups, hubs, disconnected entities, outliers, and other interpretable motifs. When the network grows larger, with hundred entities or millions of them, it becomes impossible to perform the visual analysis only at the entity level. The graph has to be summarized, and typically social scientists want to organize it in social *communities*. A large number of algorithms are available today to compute *clusters* of entities from a graph, with the assumption that the computed clusters represent faithfully the social communities. However, most social scientists are not familiar with all of the available algorithms and are challenged to choose which algorithm to run, with which parameters, and how to reconcile the computed clusters with their prior knowledge. Furthermore, the clusters computed by the algorithms do not always align with the concept of community from the social scientists.

Typically, social scientists select an analysis tool based on their familiarity with the tool and the level of local or online support they can access. Therefore, they most often use popular systems such as R [158], Gephi [13], Python with NetworkX [?], or Pajek [146]. To compute clusters, they follow a strained process: they select and run algorithms provided in the tool and then try to make sense of the results (see Figure 5.1 left). When they are not satisfied or unsure, they iteratively tweak the parameters of the algorithms at hand, run them again and hope to get results more aligned with their prior knowledge. This analysis process is unsatisfactory for three main reasons:

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

2. even when a parameter makes sense to them, such as the number of clusters to compute, k in k-means clustering, they have no clue of what value would generate good results, and are left with trial and error;

3. even if they could painstakingly evaluate the results of all clustering algorithms according to their prior knowledge, no existing system allows users to do so easily, leading users to give up and blindly accept the results of one of the first algorithms they try.

Moreover, clustering is an ill-defined problem: for one dataset, there is no ground truth, and several partitions can be considered good according to the metric chosen to evaluate the results [106]. In SNA, this means, for example, that the same social network where links represent a global notion of proximity could be clustered to find families, friend groups, or business relationships. One partition is not necessary better than another one, but depends on the purpose of the analysis. This issue increases the need for interactive tools, which let the user specify which type of partition is expected.

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

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

1. *Specify Prior Knowledge (PK)*. Users introduce their prior knowledge of the domain by defining partial clusters. The tool then runs all available clustering algorithms.
2. *Consolidate expanded PK clusters*. Users review the list of algorithms, ranked according to how well they match the prior knowledge. They compare results and consensus, then accept or ignore suggestions to expand the prior knowledge clusters
3. *Consolidate extra clusters*. The tool suggests extra clusters on unassigned nodes. The user reviews consensus on each proposed cluster, then accepts or rejects suggestions.

The output of the process is, using a direct quote from a social scientist providing feedback on the prototype: "a clustering that is supported by algorithms and validated, fully or partially, by social scientists according to their prior knowledge". According to the need to combine data mining with visualizations [179] and inspired by the idea of letting the user collaborate with the machine to reach specific goals [**?**], the proposed approach follows a user-initiated mixed-initiative [**?**] visual analytics process.

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

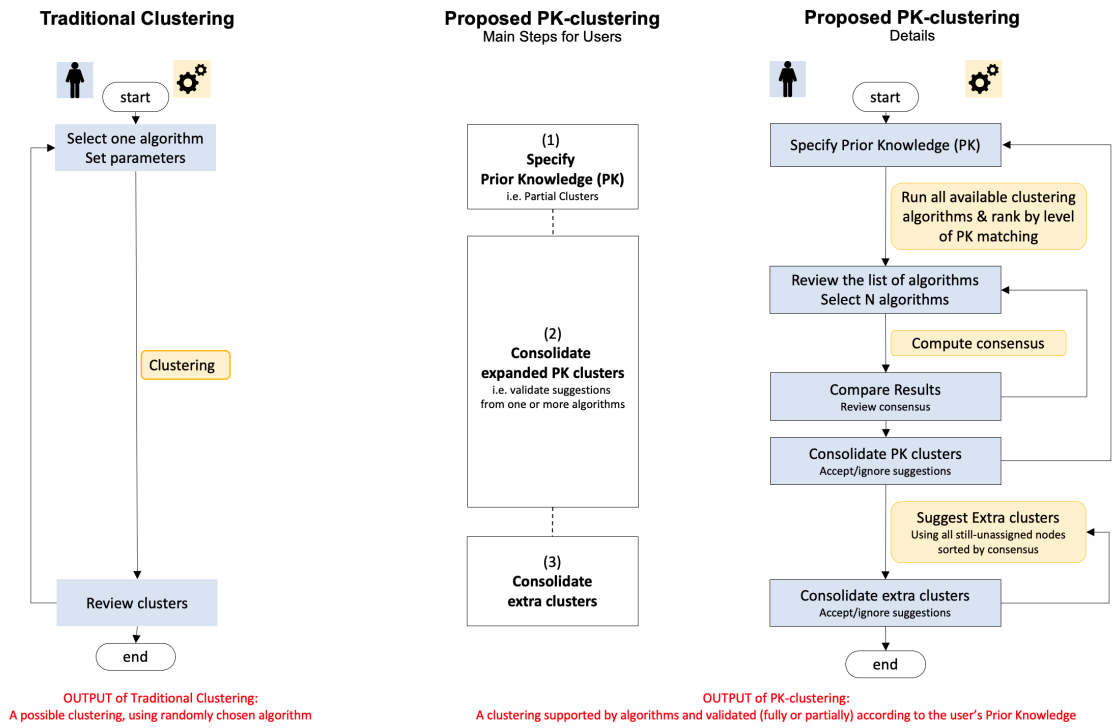

Figure 5.1 – Process of traditional clustering (left) and our PK-Clustering approach (middle and right). The output of traditional clustering is a possible clustering, using an algorithm among many choices. The output of PK-Clustering is a clustering supported by algorithms' consensus and validated (fully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

Contrary to a current trend [130], we do not aim to improve the interpretability of algorithms but to improve the interpretation of the results of black-box algorithms in light of prior knowledge, provided by the user. Every day, we use complex mechanisms that we do not fully understand, like motorbikes, cars or electric vehicles using various kinds of engines, shifts, and gears, but we are still able to choose which one best fit our needs according to their external utility and not by understanding their complex internal machinery. In addition, it is usually more important to social scientists to find an algorithm that provides useful results than to understand why another algorithm failed to do so. PK-Clustering constitute a new approach for social network clustering, that I demonstrate with a concrete prototype and validate with two case studies.

5.2 Related Work

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

5.2.1 Graph Clustering

One of the main properties of social networks is their community structure [78] that reveals group relationships between nodes, known as communities or clusters, having higher density of edges than the rest of the graph. Similar characteristics or roles are often shared between nodes of the same community. In social networks, a community can mean a lot of things like families, workgroups, or friend groups. There is abundant and growing literature on clustering methods to find these communities for social networks. The majority of the research is made only on topological algorithms, i.e., algorithms which use only the structure of the network to find clusters. [69] proposes a description and a classification of various algorithms, such as divisive, spectral, and dynamic algorithms, or methods, such as modularity-based, statistical inference, to cite a few. In contrast, many multidimensional clustering algorithms use a distance function as parameter, but graph clustering algorithms mainly rely on the structure of the graph instead.

Even if the majority of studies are based on simple graphs, real-word phenomena are often best modeled with bipartite graphs, also known as 2-mode networks. It is the case for social historians, who often build their networks from raw documents containing mentions of people, as discussed in chapter 3. Several algorithms exist for bipartite graph community detection [**?**].

Moreover, recent new approaches try to use the attributes of the nodes [**?**] and the dynamic aspect of the networks [164] to find more relevant communities. Some toolkits offer a large number of algorithms; for example, the Community Discovery Library (CDLIB) [165] implements more than 30 clustering methods with variations inspired by 67 references.

5.2.2 Semi-supervised Clustering

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

Semi-supervised clustering can be divided into constraint-based and seed-based clustering. The former includes must-link (ML) and cannot-link (CL) constraints [14, 201]. ML(x, y) indicates that given two items x and y, they must belong to the same cluster, while CL(x, y) means that x and y must belong to different clusters.

Seed-based clustering requires a small set of seeds to improve the clustering quality. Several works addressing seed-based clustering have been proposed in the literature, such as: k-means [14], Fuzzy-CMeans [**?**], hierarchical clustering [21], Density-Based Clustering [**?**], and graph-based clustering [201]. Shang et al. [**?**] use a seeding then expanding scheme to discover communities in a network. Their clustering method considers links as documents and nodes as terms.

Swant and Prabukumar [172] review graph-based semi-supervised learning methods in the domain of hyperspectral images. Vertices of the graph represent items that may be labeled, while the edges are used to specify the similarity among the items. The technique classifies unlabelled items according to the weighted distance from the labeled items.

5.2.3 Mixed-Initiative Systems and Interactive Clustering

Introduced by Horviz [**?**], mixed-initiative systems are "interfaces that enable users and intelligent agents to collaborate efficiently". Several Visual Analytics systems are based on

mixed-initiative interactions, e.g. [?, ?, 122, 212], in particular the interactive clustering systems.

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

5.2.4 Groups in Network Visualization

To assess the quality of clusters in networks, the clusters should be visualized. A state of the art report (STAR) on the visualization of group structures in graphs is proposed by Vehlow et al. [199]. Several strategies exist to display group information on top of node-link diagrams. Jianu et al. evaluated four of them: node coloring, LineSets, GMap and BubbleSets [?]. They show that BubbleSets is the best technique for tasks requiring group membership assessment. But, displaying group information on a node-link diagram can reduce the accuracy by up to 25 percent when solving network tasks. Another finding is that the use of GMap of prominent group labels improves memorability. Saket et al. evaluated the same four strategies [170], using new tasks assessing group-level understanding.

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

5.2.5 Ensemble Clustering

In the context of machine learning, an ensemble can be defined as "a system that is constructed with a set of individual models working in parallel whose outputs are combined with a decision fusion strategy to produce a single answer for a given problem" [?]. Several strategies exist for combining multiple partitions of items in a clustering setting [186]. Concerning visualization research, Kumpf et al. [?] consider ensemble visualization as a sub-field of uncertainty visualization, for which some surveys exist [?, ?]. They describe a novel interactive visual interface that shows the structural fluctuation of identified clusters, together with the discrepancy in cluster membership for specific instances and the incertitude in discovered trends of spatial locations. They aim at identifying ensemble members that can be considered similar and propose three different compact representation of clustering memberships for each member. Our system provides a consensus based interactive strategy that takes into account user's prior knowledge instead of relying on mathematically defined optimal assignments only.

5.2.6 Summary

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

Standard clustering. One algorithm is picked with a set of parameters and the user check if the results are consistent with his prior knowledge, which is not represented in the process.

Ensemble clustering. Many algorithms run with potentially many parameters, and a final partition is obtained by trying to merge optimally the partitions. At the end of the process, one clustering is given to the user who has to check if it is consistent with the prior knowledge, which is not used either.

Semi-supervised clustering. The user provides the prior knowledge and lets the algorithm propose a final solution using this information in its computation. The results should be good by design, regarding the knowledge of the user.

The aim of our proposed framework is to combine these three approaches, to integrate users in the analysis loop and allow them to have a better impact on the final community detection result.

5.3 PK-clustering

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

PK-Clustering is not tied to any specific network representation technique and could be used to augment any of them. Our prototype is implemented in the PAOHVis tool [196] which illustrates how users can view their networks as PAOH (Parallel Aggregated Ordered Hypergraph) or traditional Node Link diagrams. PK-Clustering relies heavily on having a list of nodes, so the PAOH representation is naturally adapted to PK-Clustering, and will be used in all the figures.

After a general overview of the process, I describe each step in more details, illustrated with screen samples taken during the analysis of a small fictitious network.

5.3.1 Overview

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

(1) Specify Prior Knowledge (PK).

1. The domain experts interactively specify the PK by defining groups, i.e., naming groups and assigning entities to them. Typically, an expert would assign a few items (1-3) to a few groups (2-5), thus creating a set of partial clusters.
2. All available clustering algorithms are run. Algorithm parameters (e.g., number of clusters) may also be varied manually or automatically using a grid search or a more sophisticated

strategy, resulting in additional results. Depending on the type of algorithm, topology and/or data attributes are used. The specified PK can also be used in the computation of semi-supervised clustering algorithms.

(2) Consolidate expanded PK clusters.

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

(3) Consolidate extra clusters.

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

At any point users can go back, select different algorithms, or even change the PK specification to add new partial clusters. Users can also opt not to specify any PK at all, and accept all consensual suggestions without reviewing them in details. This gives users control over how much they want to be involved in the process. Similarly, users are not required to assign every single node to a cluster, as it often happens that social scientists do not have a strong opinion on the group appartenance of some individuals. By specifying the PK in the first phase, before running the algorithms, users avoid being influenced by the first clustering results they encounter. The process leads to algorithms whose results match the PK, but it also allows to review results that contradict it.

We believe that PK-clustering addresses the important problems identified in the introduction: it helps users decide which algorithm(s) to use, facilitates the review of the results taking into consideration both the consensus between algorithms and the knowledge users have of their data. We will now review each step in more details.

5.3.2 Specification of Prior Knowledge

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

§5.2.2 in a simple visual and compact form. It is not required to specify all binary constraints because the information is derived from the prior knowledge groups.

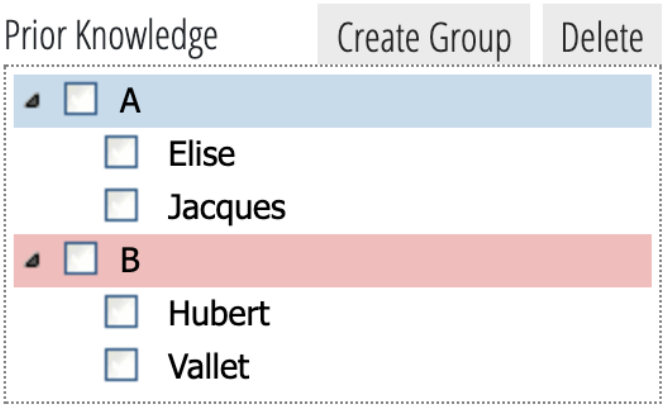

Figure 5.2 – Prior Knowledge specification, the user defined two groups composed of two members.

5.3.3 Running the Clustering Algorithms

The prototype includes 11 algorithms taken from three families:

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

Topology based algorithms. Most of the clustering algorithms consider only the graph topology and optimize a topological measure such as *modularity* [29]. Those algorithms only use the connections between the people to find groups. Their aim is to find groups of nodes such that the density of edges is higher between nodes of the same group compared to the rest of the graph.

Propagation / Learning based algorithms. Semi-supervised machine learning algorithms learn from an incomplete labeling and use it to classify the rest of the data. They represent a class of machine learning methods, also called label propagation methods, which can take into account users' PK groups in its clusters computation. By design, this type of algorithms will always provide a perfect match with the PK, even if the PK does not make much sense.

Our prototype implements 2 attribute based algorithms (one for numerical attributes and the other for categorical attributes), 9 topology based algorithms and 2 propagation based. Since we often deal with hypergraphs, 2 of the topology-based algorithms are bipartite node clustering algorithms: Spectral-co-Clustering [?] and Bipartite Modularity Optimisation [?]. Since the majority of community detection algorithms are for unipartite graphs, the system performs a projection into a one-mode network [?]. Basically, each pair of nodes which are

in the same hyperedge are connected together in the resulting graph, with a weight being the number of shared hyperedges [?].

Some algorithms require parameters to be specified. We do not force the user to specify values for all the parameters, when possible, we infer them from the PK-groups. For instance, instead of using an arbitrary default for the number of expected clusters k in k-means clustering, we run the algorithm several times with a value of k from the number of specified PK-groups to this number plus two. The strategy of using several parameter combinations for the same algorithm is often used in ensemble clustering to increase the number of different clusterings. However, the number of parameter combinations can be extremely high. The research field of *visual parameter space exploration* (see e.g., [?]) is devoted to exploring this space of parameter values in a sensible way; we currently address the problem only for simple cases.

Once all the algorithms finish the computation, the resulting clustering are matched with the PK and ranked by how interesting their results are likely to be for the user.

5.3.4 Matching Clustering Results and Prior Knowledge

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

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

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

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

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

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

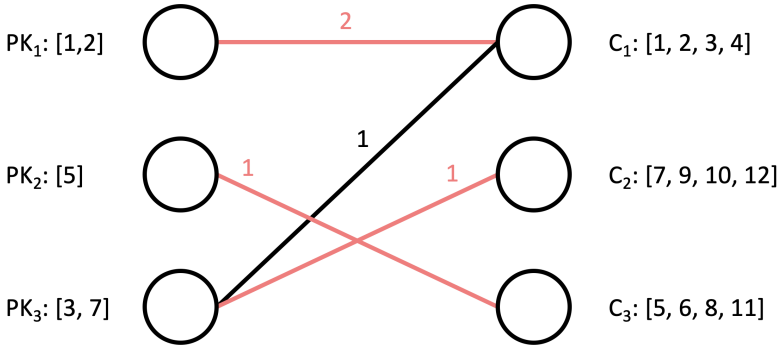

Figure 5.3 – Red edges represent the prior knowledge matching

For example, given a clustering of 12 nodes $N = 1, 2, \ldots, 12$, the clusters $C_1 = [1, 2, 3, 4]$, $C_2 = [7, 9, 10, 12]$ and $C_3 = [5, 6, 8, 11]$ and a PK composed of 3 groups $PK_1 = [1, 2]$, $PK_2 = [5]$ and $PK_3 = [3, 7]$, the maximum-weight matching is given by the edges (PK_1, C_1), (PK_2, C_3) and (PK_3, C_2). This is illustrated in Figure 5.3. The edges of the matching correspond to the matching between the PK-groups and the clusters. The edit distance is then equal to the sum of all the weights of the bipartite graph minus the sum of the weights of the maximum matching (in red), thus equaling $5 - 4 = 1$. In other words, we only have to move the node 3 from PK_3 to PK_1, for every PK-group to be a subset of an unique cluster, with no overlap.

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

5.3.5 Ranking the Algorithms

The algorithms are ranked by their degree of matching with the PK, using the edit distance. We also introduce a *parsimony* criterion if there is a tie between two or more algorithms. The algorithm with the smaller number of other clusters will be shown first, as the results are easier to interpret. Moreover, the number of specified PK groups is expected to be close to the final number of clusters the user wants to retrieve, as social scientists often have a good knowledge of their data.

To complement the parsimony rule, we also consider that the family of propagation/learning based clustering algorithms is more complex than the two previous families (attribute or topological based clustering), in the sense that they are more difficult to explain. If a simple and a

complex algorithm match the prior knowledge, the simpler one is presented first. For example, if grouping by the attribute "profession" provides a perfect match, then it is ranked higher than a propagation based method achieving the same perfect match.

Semi-supervised methods will always provide a perfect match by definition. But if all the other algorithms (topological and attribute based) do not give a match, it means that the PK does not align well with the data. This would signal users to reconsider their PK, as it does not match the data encoded in the network.

5.3.6 Reviewing the Ranked List of Algorithms

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

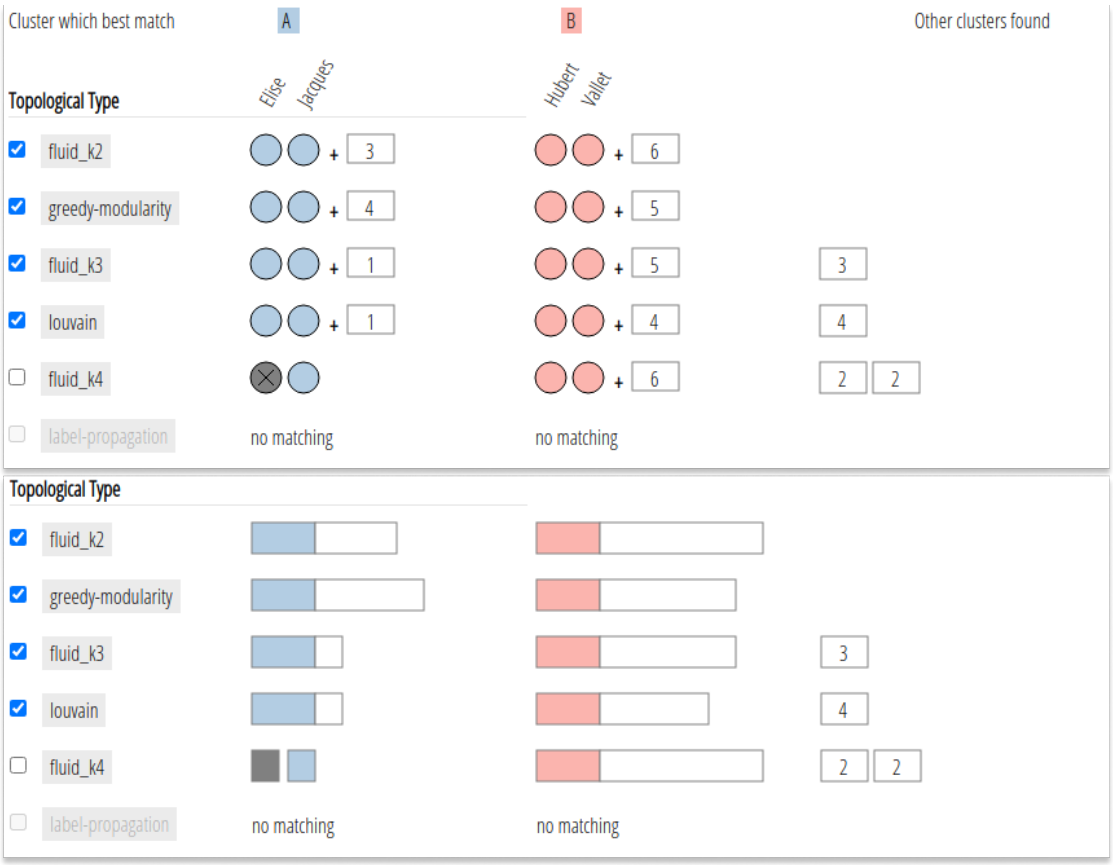

Figure 5.4 – Two different modalities for the ranked list of algorithms. Top: persons are shown as circles. Bottom: aggregated view. Colors indicate the matching group. Gray indicates no match. White indicates extra nodes or clusters.

Each row is an algorithm, and the algorithms are grouped by family. On the right of the name of the algorithm there is a representation of the clusters that best match each of the

PK-groups. Figure 5.4 shows first the cluster which best matches the blue PK-group, and then the cluster which best matches the red PK-group. In each cluster we see colored dots for each person that matches, and dark gray dots with a X for no match. Additional nodes in the cluster are represented as white dots with a number next to it. On the right most we see how many other clusters (if any) have been found by the algorithm—also represented as white dots with a number next to it.

For example, the second algorithm *fluid_k3* has a blue cluster that matches the blue PK-group plus 1 extra node, a red cluster that matches the red PK-group plus 5 nodes, and one extra cluster. The top four algorithms match the PK perfectly, while the following one *fluid_k4* has a partial match. At the bottom, an algorithm (*label-propagation*) has no match.

The alternate modality of representing the matches (shown at the bottom of Figure 5.4) uses bars to aggregate the nodes and show the proportion of matching, non-matching and other nodes in each cluster. This is more useful when dealing with larger networks, because it allows users to see the results in a more compact way.

Once users have reviewed the list of algorithms they can review results of a single algorithm, or review and compare the results of all the selected algorithms. By default only the top algorithms are selected for inspection, but users can select any set of algorithms according to different criterion: the *degree of matching* (i.e., they can choose to look at algorithms with no match to challenge their prior knowledge); the *algorithm type* (the user may prefer an attribute-based algorithm, rather than one based on topology); the *size* of the matched clusters; or the number and size of *other clusters* found by the algorithm.

PK-Clustering expresses its prior knowledge through *must-link* and *cannot-link* constraints. However, at this stage, the user can decide to use this expressive power as strong constraints—only selecting algorithms that match all of them—or as weak constraints—to explore clustering results that support most or some of them. Our historian colleagues have used both, either to cluster a well-understood dataset with strong constraints or to generate hypotheses on less known ones.

5.3.7 Reviewing and Consolidating Final Results

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

Reviewing Results of a Single Algorithm

By clicking on an algorithm name the results of that algorithm are displayed in the PAOHVis view (see Figure 5.5). In this view, each line corresponds to a person in the graph, and each vertical line represents an hyperedge connecting them [196], in a way visually similar to the UpSet

representation [?] but semantically different. Alternative graph representations are available as well—such as node link diagrams—but the PAOHVis view is well adapted to PK-Clustering.

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

Comparing Multiple Algorithm Results

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

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

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

In Figure 5.5 PK_louvain is selected as the base algorithm for the grouping of names in the list. We see that there is very good consensus on the red cluster, but in the blue cluster only 4 out of 7 algorithms see Joseph as belonging to it. Others see him as belonging to the red cluster. In *Others*, 4 algorithms consistently disagree by assigning 3 more nodes to the blue cluster. There are clearly many ways to cluster data, and users must decide the more meaningful one, based on their deep knowledge of the people in the network before validating clusters, possibly by re-reading source documents or gathering more.

Consolidating the prior knowledge clusters

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

Partial Copy. By clicking on one of the icons or dragging the cursor down on a set of icons, users validate the suggestion(s) of an algorithm, adding colored squares in the consolidation

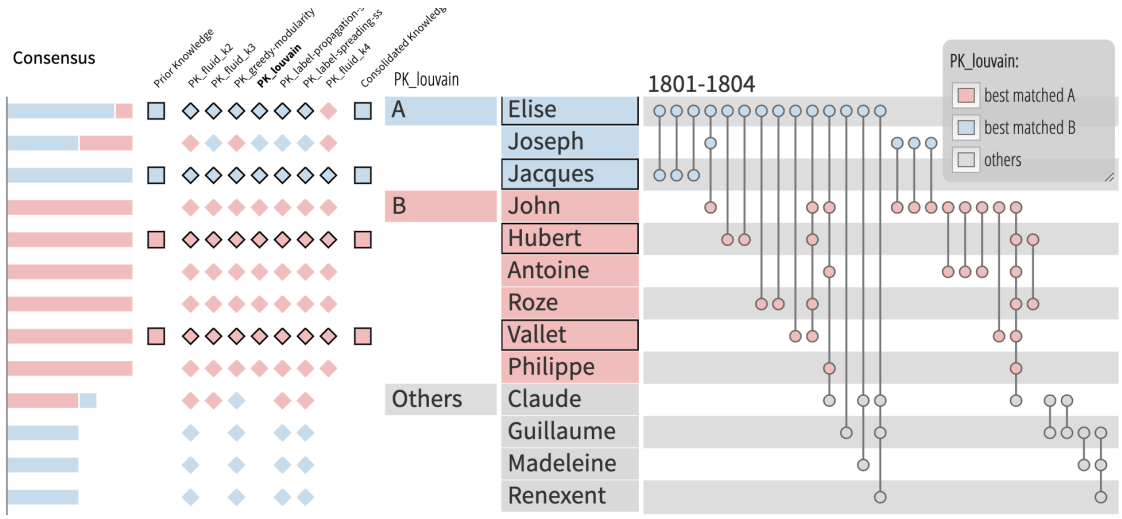

Figure 5.5 – Reviewing and comparing results of multiple algorithms. One algorithm is selected to order the names and group them, but icons show how other algorithms cluster the nodes differently, summarized in the consensus bar on the left.

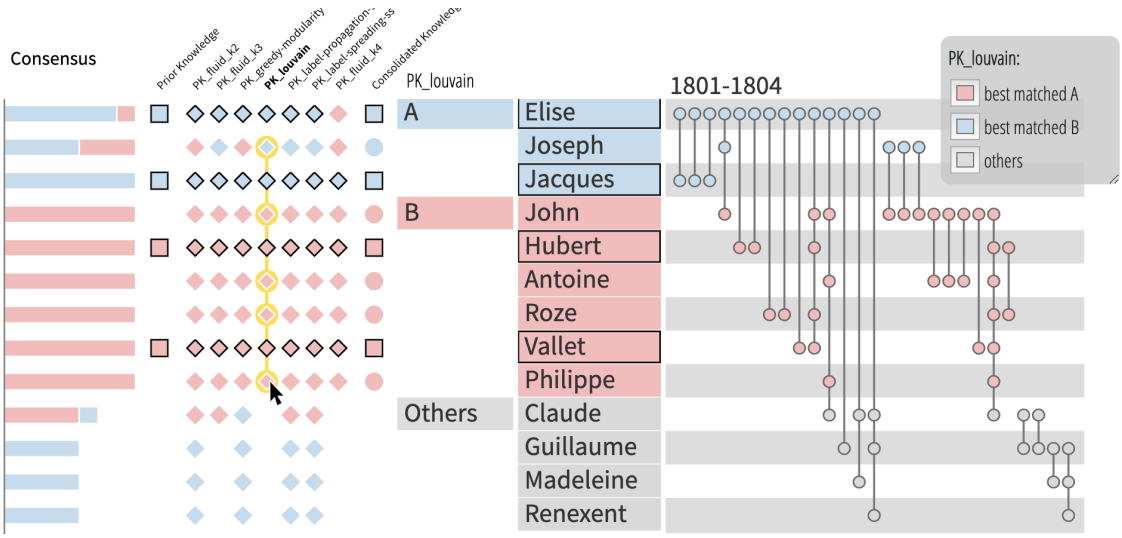

Figure 5.6 – The user quickly drags on consecutive icons (in yellow) representing the suggestions made by one algorithm to validate node clustering. Once the cursor is released the validated nodes appear as squares icons in the Consolidated Knowledge column.

column. Once this validation is done, the squares do not change color anymore and represent the user's final decision (unless changed manually again). Figure 5.6 shows how a user drag-selects a set of diamonds in the column PK_fluid_k4. They are connected by a yellow line, which appears while dragging over the icons. When done the status of the nodes in the Consolidated Knowledge column (rightmost) will change to square.

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

In summary, diamonds represent suggestions from one algorithm, circles temporary choices, and squares represents the knowledge validated by the user.

Direct tagging. At any time, users can manually overwrite the association of a node to a cluster by right clicking on the node in the consolidated knowledge column and selecting an cluster from a menu. When no clear decision can be made users can leave nodes unassigned, and no shape is displayed in the consolidated knowledge column.

Consolidating extra clusters

The last step of PK-clustering aims to find new clusters for the nodes that have not been validated yet, based on the consensus of the selected algorithms. The suggestions are made from the point of view of one clustering algorithm that users can change along the process. First, the user selects one algorithm in the PAOHVis view and the nodes are grouped by the clusters found by the algorithm. The PK-clusters are displayed at the top, followed by *Others*, which contains everyone else. When users click on *Others*, the other clusters are displayed ordered by consensus. Since the number of clusters can be high, all new clusters appear in gray to avoid the rainbow effect. A secondary matching process matches the clusters of the current algorithm with those of all the other algorithms, one by one (similar to the matching process described in §5.3.4) . Once the matching is done, the consensus of one cluster is computed as the sum of the cardinalities of the intersections between the cluster and all the other clusters of the other algorithms matched with it, divided by the number of nodes of the cluster.

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

The evaluation of the best cluster for a node can be done using multiple encodings. The suggested clusters appear into the consensus bar chart, in the set of algorithm output and when hovering over the node. A click on the color will validate the node into the cluster having that color. If users are satisfied with the association proposed by the current algorithm, they can validate it by clicking on the cluster name. This will create a new group, so the user can classify the nodes into this new group, as seen before (§5.3.7): using the consensus slider, copying an

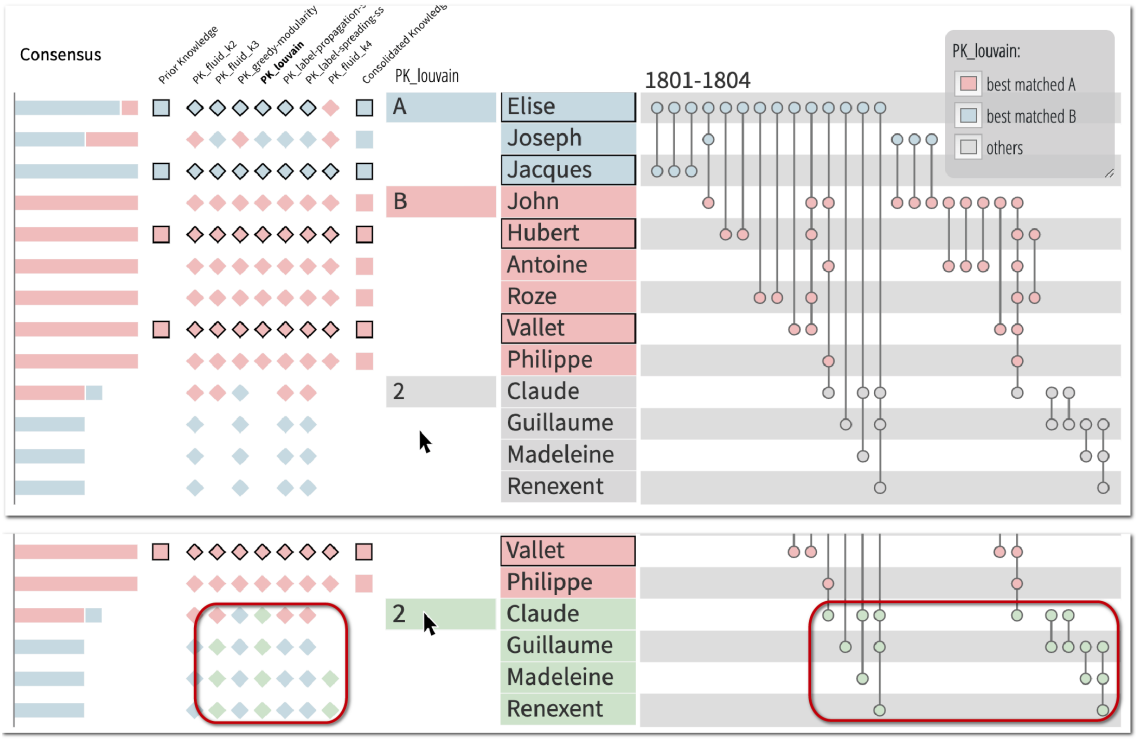

Figure 5.7 – Suggestion of extra clusters. The two PK-groups (red and blue) are validated (nodes in the consensus column are all squared). One extra clusters is proposed by the Louvain algorithm, labeled as 2. Hovering over the cluster 2, the consensus is displayed by the green diamonds. This feedback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

algorithm result, or through manual labeling. This process is repeated for the other clusters until there are no unlabeled nodes or the user is satisfied with the partial clustering. An example of a fully consolidated dataset is shown in Figure 5.8.

5.3.8 Wrapping up and Reporting Results

At any stage of the process, the user can finish instantaneously, either by not labeling undecided nodes, or selecting and validating the results of a single algorithm—as traditional approaches do, or by using a specified threshold of consensus and not labeling the remaining entities. The appropriateness of the choice is up to the user and should be documented in the publication.

In addition to the consolidated clustering, the output of PK-clustering consists of provenance information in the form of a table and a summary report. The table provides, for each node, the consolidated label, along with the labels produced by all the selected algorithms, and a description of the interaction that has led to the consolidation, such as "selected from `algorithm x`", "consensus >= 5", or "override" when manually selected by the user instead of selected from an algorithm. The summary provides counts of how many nodes were labeled

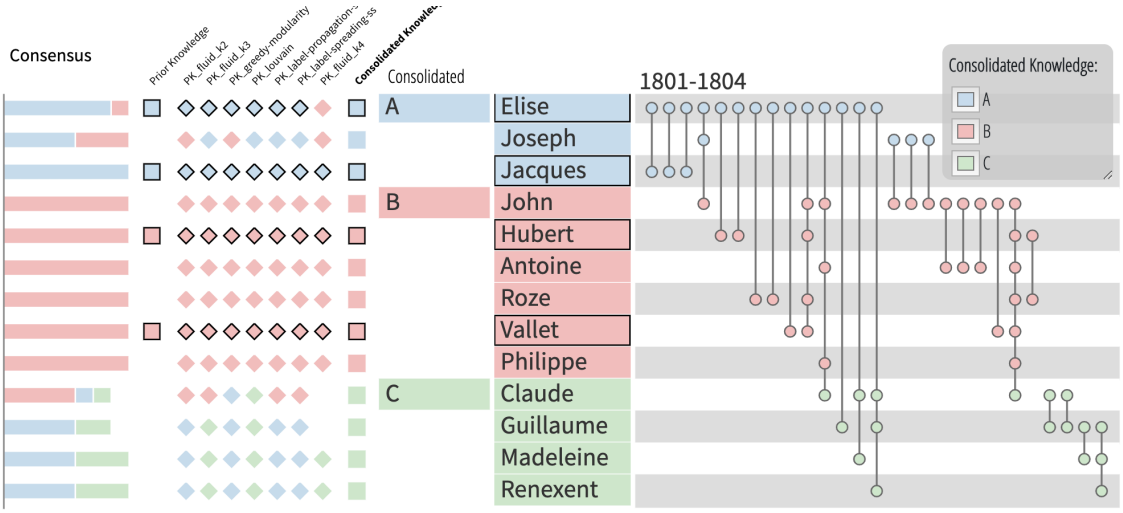

Figure 5.8 – The dataset has been fully consolidated. The persons are grouped and colored by the consolidated knowledge. The user decided to assign Claude, Guillaume, Madeleine and Renexent to cluster *C*, by taking into account the graph and the consensus of the algorithms.

using the different interactions methods and can be used in a publication.

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

5.4 Case studies

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

5.4.1 Marie Boucher Social Network

I asked one of our historian colleague her prior knowledge on her network about the trades of Marie Boucher [?], composed of two main families: Antheaume and Boucher. Family ties were important for merchants, but could not scale above a certain level. Marie Boucher expanded her trade network far beyond that limit. She then had to connect to bankers, investors, and foreign traders, far outside her family and yet connected to it indirectly. As hinted in her article, Dufournaud believes that the network can be split in three clusters: one related to the Boucher family, one to the Antheaume family, and the third to the Boucher & Antheaume company. Using standard visualization tools, she could see different connection patterns over time, but she wanted to validate her hypothesis using more formal measures and computational methods.

So she specified her hypotheses as PK and started the analysis. Figure 5.9 (top left) shows the three PK groups: Marie Boucher for the Boucher family, Hubert Antheaume for the

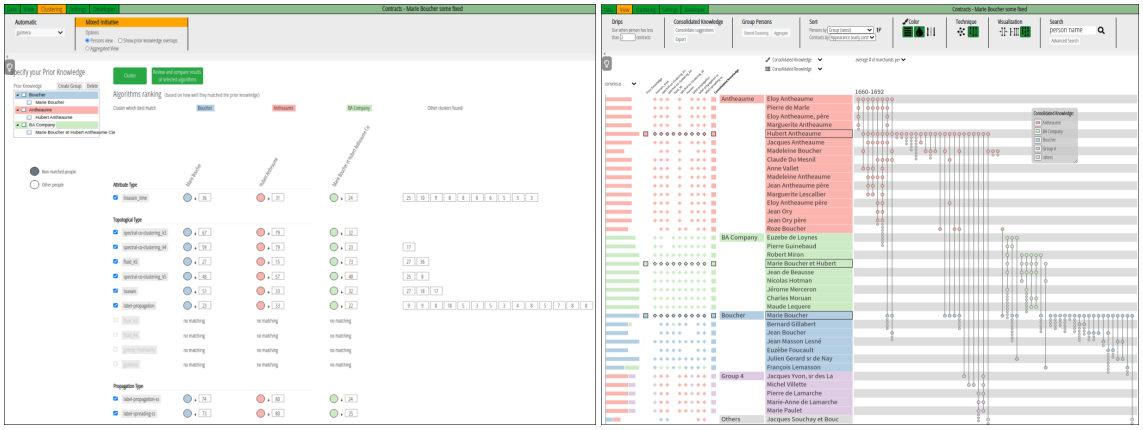

Figure 5.9 – Two main phases of PK-clustering. On the left, the user has specified the Prior Knowledge (PK) groups (top left) and then reviews the list of algorithms ranked according to how well they match the PK. On the right, the user compared the detailed results of selected algorithms and consolidated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

Antheaume family, and the Boucher & Antheaume corporation alone for the company.

After running the algorithms, 9 algorithms produced a perfect match out of the 13 executed (see Figure 5.9 - left.) with the first algorithm listed an attribute based algorithm that uses the time attribute in its computation. That summary alone was found very interesting because the 3 clusters seemed very consensual among all the 9 algorithms, and furthermore, they appeared explainable by time alone.

In the PAOH view, she started by consolidating the 3 PK-groups using the amount of consensus among the algorithms as well as the network visualization and her own knowledge of the persons. At the end of this step, the Boucher, Antheaume, and Boucher & Antheaume groups were consolidated, but there were still several persons not labeled on the consolidated knowledge. She decided to review in more detail the clustering results using the *ilouvain_time* algorithm because of its reliance on the time attribute, and also because its results seemed good in the matching view. After clicking on the virtual group *Others*, the four other clusters computed by *ilouvain_time* appeared and were reviewed by hovering the mouse on the names of these new groups. She selected only one clusters she was confident about and consolidated it.

The final validated partition of the dataset is represented in Figure 5.9 (right). The persons are colored and grouped by the consolidated knowledge. We can see that the final grouping makes sense in the PAOH visualization on the right. Only one person is not part of any group: Jacques Souchay. It is not unusual in historical sources to have persons mentioned without any information on them.

Our historian colleague can now publish a follow-up article validating her hypotheses. The summary report will help document where the final grouping came from, increasing trust with

Figure 5.10 – Computing the Lineages of VAST authors: Prior Knowledge from Alice and results of the clusterings matching it.

regard to her claims.

5.4.2 Lineages at VAST

In the second cases study we took the role of Alice, a VAST Steering Committee (SC) member, who participates in a SC meeting to validate the Program Committee proposed by the VAST paper chairs for the next conference. One of the many problems that all conference organizers face is to balance the members of the Program Committee according to several criteria. The InfoVis Steering Committee Policies FAQ states that the composition of the Program Committee should consider explicitly how to achieve an appropriate and diverse mix [97] of: • academic lineages • research topics • job (academia, industry) • geography (in rough proportion to the research activity in major regions) • gender. Most of these criteria are well understood, except *academic lineage* which is not clearly defined. Alice will use the "Visualization Publications Data" (VisPubData [?]) to find-out if she can objectify this concept of lineage to check the diversity of the proposed Program Committee accordingly.

Using PK-clustering, Alice loads the VisPubData, filtered to only contain articles from the VAST conference, between 2009–2018. Only prolific authors can be members of the program committee, but highly filtering the co-authorship network would change its structure and disconnect it. Thus, she will use the unfiltered network of 1383 authors to run the algorithms and perform the matching (Step 1 of the process), even if at the end only 113 authors with more than 4 articles will be consolidated (Steps 2 and 3).

Alice starts the PK-Clustering process by entering her prior knowledge, which is partial and based on two strategies: her knowledge of some areas of VAST, and the name of well-known researchers who have developed their own lineage. She runs the algorithms (Figure 5.10) and 5 algorithms produce a perfect match, acknowledging her knowledge of some areas of VAST. She then shows the results to other members of the SC who will help her consolidate the lineage clusters.

Her initial PK clusters are quickly consolidated, using Internet search to validate some less known authors. She then decides to create as many additional clusters and lineage groups

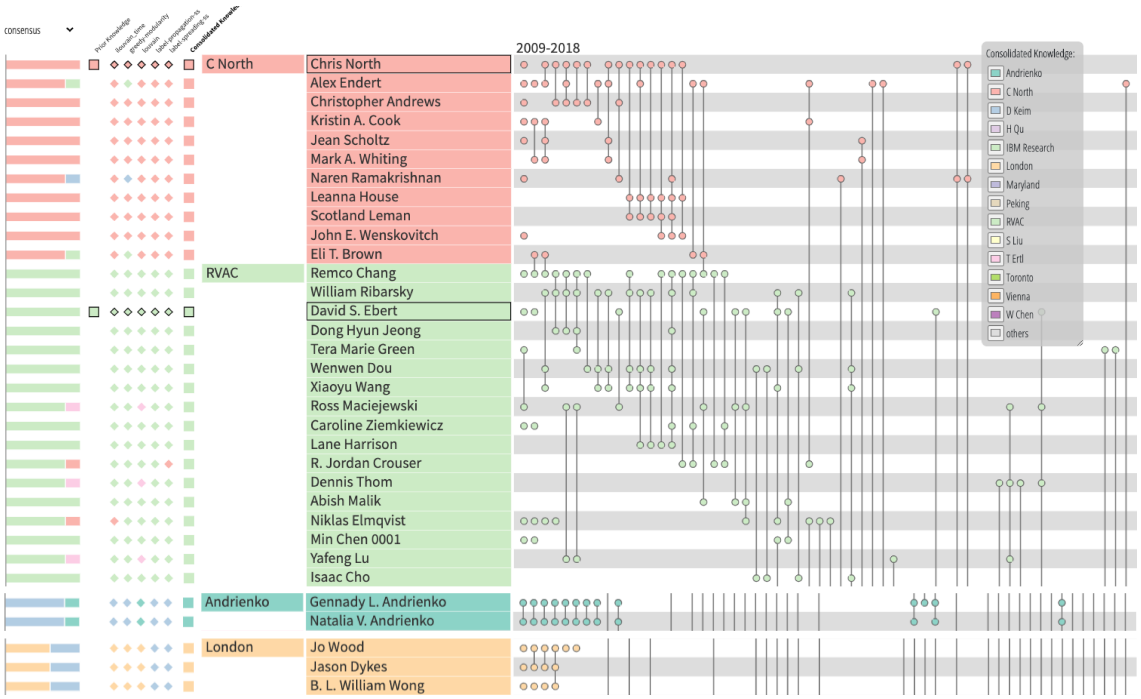

Figure 5.11 – Four consolidated groups in the VAST dataset: C North, RVAC, Andrienko and London

as she can. For some authors, she decides to override the consensus of the algorithms. For example, she decides, and her colleagues agree, that Gennady and Natalia Andrienko should be in their own lineage group and not in D. Keim's (Figure 5.11). The history of VAST in Europe, very much centered around D. Keim and the VisMaster project [200], has strongly influenced the network structure and some external knowledge is required to untangle it.

Using the *PK_louvain* algorithm as starting point, Alice creates new groups and achieves a consensus among the experts on a plausible set of lineages for VAST. She then checks with the list proposed by the program committee by entering it in on a spreadsheet with the names and affiliations. She adds the groups and their color, and sort the list by group. Alice can now report her work to the whole steering committee, which can check the balance of lineages according to this analysis, and decide if some lineage groups are over or under represented. By keeping the affiliations in the list, the SC can also check the balance of affiliations that is not always aligned with the lineages. The final results are available in the supplemental material of the article.

Using partitioning clustering (although with outliers) forces the algorithms or experts to make strong decisions related to lineages. But using a soft clustering (or overlapping partitions), while providing a more nuanced view of lineages, would not be as simple to interpret as coloring spreadsheet lines and sorting them; in the end, the final selection only uses the lineage criterion among many others. Still, we believe PK-clustering can provide a partial but concrete answer to the problem of defining what the scientific lineages are.

5.4.3 Feedback from practitioners

Although we could not conduct face to face meetings with historians and sociologists due to the COVID19 lockdown, we showed the system to three practitioners and asked their feedback through videoconferencing systems, sharing video demonstrations and sharing our screen.

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

They were intrigued by the idea of entering the prior knowledge to the system, but acknowledged that it was easy to understand and natural for them to think in terms of well-known entities belonging to groups. They felt uneasy thinking that this prior knowledge could bias the results of the clustering and of the analysis. However, after a short discussion, they also agreed that the traditional process of picking in a more or less informed way two or three algorithms to perform a clustering was also probably priming them and adding other biases. Still, they said that they would need to explain the process clearly in their publications and that some reviewers could also stress the risks.

They all agreed that the process was clear and made sense, but they also felt it was complicated and that they would need time to master it. They said that it was more complicated than pressing a button, but that the extra work was worth it.

One historian who spends a lot of time analyzing her social networks and finding information about all the people was shocked by the idea that you could want to use an algorithm that did not match fully the prior knowledge. For us, it matters if the prior knowledge is given as constraints or preferences, but we did not want to introduce these notions in the user interface so analysts are free to interpret the prior knowledge as one or the other.

They also identified some issues with the prototype. It was not managing disconnected networks at all when we showed the demo, and they stressed the fact that real networks always have disconnected components. They were also asking about structural transformations, such as filtering by attribute or by node type. We chose not support these functions at this stage, but they can be done through other standard network systems.

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

The table and summary report was added after those sessions so no feedback was gathered. We will continue to collaborate with those practitioners and help them test PK-culstering during their next social network analysis project.

5.5 Discussion

As presented in §5.2.6, the existing approaches to create clusters in social networks consider three options: standard clustering, ensemble clustering, and semi-supervised clustering. The proposed PK-Clustering approach combines aspects of the three options in order to give more control to users in the analysis loop, and allow them to have more say in the final results.

Proponents of automatic methods may argue that PK-clustering gives users too much influence on the final result as they can change the cluster assignments at will. On the other hand, social scientists are rarely satisfied with current clustering methods, in part because they run on network data that rarely represent all the knowledge they have of the social network, so providing user control to correct mistakes is critical.

Traditional methods push users to believe the results of the first algorithms and parameter selection they try (typically chosen randomly). Using PK-Clustering, users can still follow blindly the results of one algorithm if they want but the system provides a more systematic approach. It allows users to compare results, review consensus, think at each phase and reflect on decisions. Instead of passively accepting what the algorithms propose, users provide initial hypotheses—which limits the chances of being primed by an algorithm, and explicitly validate the cluster assignment of nodes, therefore performing a critical review of the automated results, yet with fast interaction to accept many suggestions at once when appropriate.

This new approach allows users to discover alternative views. For example when algorithms do not match the PK, it is an indication that the PK is being challenged and may not be correct. Users actively participate in the process of assigning, a requirement for social scientists. The report produced at the end of the analysis adds transparency by recording where the results come from for each node so decisions can be reviewed. Ultimately social scientists remain responsible for reporting and justifying their choices and interventions in their publication.

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

5.5.1 Limitations

Many more clustering algorithms exist and could be added. Moreover, expanding the exploration of parameter spaces for clustering algorithms seems needed. Another limitation of the current prototype is that some algorithms do not work well with disconnected components of the graph. Unfortunately, social scientists datasets typically have many disconnected components. This issue can be mitigated by separating components into a set of connected components, run the algorithms on them, and merge the results. The prototype runs both with node-link and PAOH representations, but it is better tuned to the PAOH representation because of its highly readable nodes list and table format which makes the review of consensus easier. Better coordination of the table with node-link diagrams and other network visualizations is needed. Further case studies could also help us improve the utility of the tool as well as the provenance table and summary.

5.5.2 Performance

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

In order to improve the computational scalability, progressive techniques can help to deal with larger sizes [**?**]. The current user interface design for PK-Clustering would allow the ranked list of algorithms to be progressively updated, and users to review a few individual algorithms first while other algorithms are still running. Of course, visual scalability is also an issue with larger datasets, as the list of people also grows. PAOHVis allows groups (like clusters) to be aggregated or expanded, so we expect that users would expand clusters one by one to review and consolidate them, while also being able to review the connections between the proposed clusters. Users can also use the automated features of PK-Clustering to consolidate the nodes (e.g., selecting one algorithm based on the ranking, or using the consensus slider to consolidate all the nodes at once). Pixel-oriented visualizations [**?**] would facilitate the review of consensus for a large number nodes and clusters. Classic techniques like zooming or fisheye views [**?**, 160] would help as long as names remain readable, which is critical to our users.

5.6 Conclusion

In this chapter, I introduced a new approach, called PK-Clustering, to help social scientists create meaningful clusters in social networks. It is composed of three phases: 1) users specify the prior knowledge by associating a subset of nodes to groups, 2) all algorithms are run and ranked, 3) users review and compare results to consolidate the final clusters.

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

PK-Clustering follows traceability, simplicity, and document reality properties discussed in

chapter 1 and chapter 3, by respectively providing a summary report of the actions leading to the final clustering, simple interactions, and the usage of bipartite multivariate dynamic networks as a data model. This approach is a concrete proof-of-concept solution to **Q3** in the context of clustering, as it provides a framework for social scientists, specifically historians, to follow a clustering analysis supported by algorithmic power but always in control of the decision process, through easy-to-use interactions. Clustering and social network analysis remain a challenging task, typically without ground truth to formally evaluate the results. If PK-Clustering limits bias inherent to traditional clustering (priming bias and lack of control), the high influence of users on the decision-making may introduce other type of bias. Still, I believe that PK-Clustering offers a fresh perspective on the process of clustering social networks and gives users the opportunity to report their results in a transparent manner.

Bibliography

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

[2] Moataz Abdelaal, Nathan D. Schiele, Katrin Angerbauer, Kuno Kurzhals, Michael Sedlmair, and Daniel Weiskopf. Comparative Evaluation of Bipartite, Node-Link, and Matrix-Based Network Representations, August 2022. arXiv:2208.04458. 28

[3] Ruth Ahnert, Sebastian E. Ahnert, Catherine Nicole Coleman, and Scott B. Weingart. The Network Turn: Changing Perspectives in the Humanities. *Elements in Publishing and Book Culture*, December 2020. doi:10.1017/9781108866804. 19

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

[5] Mashael Alkadi, Vanessa Serrano, James Scott-Brown, Catherine Plaisant, Jean-Daniel Fekete, Uta Hinrichs, and Benjamin Bach. Understanding barriers to network exploration with visualization: A report from the trenches. *IEEE Transactions on Visualization and Computer Graphics*, 27(2), 2022. 2, 7, 9, 12, 26, 29, 33, 34, 36, 40, 42, 48, 107

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

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

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

[9] Mariona Coll Ardanuy, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel CS Wilson, and Barbara McGillivray. Living Machines: A study of atypical animacy, November 2020. arXiv:2005.11140, doi:10.48550/arXiv.2005.11140. 19

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

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

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

[13] 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, 29, 53, 61, 84

[14] Sugato Basu, Ian Davidson, and Kiri Wagstaff. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC, first edition, 2008. 87

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

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

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

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

[19] A. Bezerianos, F. Chevalier, P. Dragicevic, N. Elmqvist, and J.d. Fekete. GraphDice: A System for Exploring Multivariate Social Networks. *Computer Graphics Forum*, 29(3):863–872, 2010. doi:10.1111/j.1467-8659.2009.01687.x. 61

[20] Marc Bloch. *Apologie Pour l'histoire*. A. Colin, 1949. 2

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

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

[23] Stephen Borgatti. Social Network Analysis, Two-Mode Concepts in. *Computational Complexity: Theory, Techniques, and Applications*, January 2009. doi:10.1007/978-0-387-30440-3_491. 26, 45

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

[25] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, December 2011. doi:10.1109/TVCG.2011.185. 61, 71

[26] Romain Boulet, Bertrand Jouve, Fabrice Rossi, and Nathalie Villa. Batch kernel SOM and related Laplacian methods for social network analysis. *Neurocomputing*, 71(7):1257–1273, March 2008. doi:10.1016/j.neucom.2007.12.026. xi, 8, 26

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

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

[29] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On Modularity Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188, February 2008. doi:10.1109/TKDE.2007.190689. 40, 91

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

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

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

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

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

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

[36] Raphaël Charbey and Christophe Prieur. Stars, holes, or paths across your Facebook friends: A graphlet-based characterization of many networks. *Network Science*, 7(4):476–497, December 2019. doi:10.1017/nws.2019.20. 23, 24, 54

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

[38] J. S. Coleman. Introduction to mathematical sociology. *Introduction to mathematical sociology.*, 1964. 22

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

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

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

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

[43] Pascal Cristofoli. Principes et usages des dessins de réseaux en SHS. *La visualisation des données en histoire*, page 35, 2015. 2, 28, 61

[44] Pascal Cristofoli and Nicoletta Rolla. Temporalités à l'œuvre dans les chantiers du bâtiment. *Temporalités. Revue de sciences sociales et humaines*, (27), June 2018. doi:10.4000/temporalites.4456. xii, 37, 43, 56, 60

[45] Tarik Crnovrsanin, Chris W. Muelder, Robert Faris, Diane Felmlee, and Kwan-Liu Ma. Visualization techniques for categorical analysis of social networks with multiple edge sets. *Social Networks*, 37:56–64, 2014. doi:10.1016/j.socnet.2013.12.002. 43

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

[47] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. 55, 61

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

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

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

[51] Mandeep K. Dhami, Ian K. Belton, and David R. Mandel. The "analysis of competing hypotheses" in intelligence analysis. *Applied Cognitive Psychology*, 33(6):1080–1090, 2019. doi:10.1002/acp.3550. 3

[52] 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. 7, 31, 34, 38, 39

[53] Dana Diminescu. The migration of ethnic germans from romania to west germany: Insights from the archives of the former communist regime. In *CERS, Public Lecture, UCLA*, Los Angeles, United States, March 2020. 37, 58

[54] Nicole Dufournaud. La recherche empirique en histoire à l'ère numérique. *Gazette des archives*, 240(4):397–407, 2015. doi:10.3406/gazar.2015.5321. 1, 109

[55] 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. xi, 4, 7, 37, 42

[56] Nicole Dufournaud and Jean-Daniel Fekete. Comparaison d'outils pour la visualisation de sources historiques codées en XML/TEI. *Document numérique*, 9(2):37–56, April 2006. doi:10.3166/dn.9.2.37-56. 39

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

[58] Dan Edelstein, Paula Findlen, Giovanna Ceserani, Caroline Winterer, and Nicole Coleman. Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project. *The American Historical Review*, 122(2):400–424, April 2017. doi:10.1093/ahr/122.2.400. xi, 19, 20, 33, 108

[59] P. Erdös and A. Rényi. On the evolution of random graphs. In *On the Evolution of Random Graphs*, pages 38–82. Princeton University Press, October 2011. doi:10.1515/9781400841356.38. 21

[60] Emily Erikson and Peter Bearman. Malfeasance and the Foundations for Global Trade: The Structure of English Trade in the East Indies, 1601–1833. *American Journal of Sociology*, 112(1):195–230, July 2006. doi:10.1086/502694. 43

[61] Michael Eve. Deux traditions d'analyse des reseaux sociaux. *Réseaux*, 115(5):183–212, 2002. 9, 23, 24

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

[63] Lucien Febvre. VERS UNE AUTRE HISTOIRE. *Revue de Métaphysique et de Morale*, 54(3/4):225–247, 1949. 17

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

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

[66] Robert Fogel. *Railroads and American Economic Growth: Essays in Econometric History*. 1964. 18

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

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

[69] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, February 2010. doi:10.1016/j.physrep.2009.11.002. 23, 87

[70] L. Freeman. Visualizing Social Networks. *J. Soc. Struct.*, 2000. 2, 4, 5, 27

[71] L.C. Freeman. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press, 2004. 1, 3, 11, 20, 21, 22, 23, 40

[72] Manuel Freire, Catherine Plaisant, Ben Shneiderman, and Jen Golbeck. ManyNets: An interface for multiple network analysis and visualization. In *CHI '10*, CHI '10, pages 213–222, New York, NY, USA, 2010. ACM. doi:10.1145/1753326.1753358. 55

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

[74] Michael Friendly. A Brief History of Data Visualization. In Chun-houh Chen, Wolfgang Härdle, and Antony Unwin, editors, *Handbook of Data Visualization*, Springer Handbooks Comp.Statistics, pages 15–56. Springer, Berlin, Heidelberg, 2008. doi:10.1007/978-3-540-33037-0_2. 14

[75] GEDCOM: The genealogy data standard. 27

[76] 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. 28

[77] Carlo Ginzburg and Carlo Poni. La micro-histoire. *Le Débat*, 17(10):133, 1981. doi:10.3917/deba.017.0133. 3, 24, 36

[78] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, June 2002. doi:10.1073/pnas.122653799. 23, 87

[79] Barney G. Glaser and Anselm L. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Transaction, New Brunswick, 5. paperback print edition, 2010. 3

[80] Michael Gleicher. Considerations for visualizing comparison. *IEEE Trans. Vis. Comput. Graphics*, 24(1):413–423, 2018. doi:10.1109/TVCG.2017.2744199. 55

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

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

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

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

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

[86] Klaus Hamberger, Cyril Grange, Michael Houseman, and Christian Momon. Scanning for patterns of relationship: Analyzing kinship and marriage networks with Puck 2.0. *The History of the Family*, 19(4):564–596, October 2014. doi:10.1080/1081602X.2014.892436. 25, 27, 43, 45

[87] Klaus Hamberger, Michael Houseman, and R. White, Douglas. Kinship network analysis. In John Scott & Peter J. Carrington, editor, *The Sage Handbook of Social Network Analysis*, pages 533–549. Sage Publications, 2011. 1, 27

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

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

[90] Thomas N. Headland, Kenneth L. Pike, and Marvin Harris, editors. *Emics and Etics: The Insider/Outsider Debate*. Emics and Etics: The Insider/Outsider Debate. Sage Publications, Inc, Thousand Oaks, CA, US, 1990. 42, 109

[91] J. Heer and D. Boyd. Vizster: Visualizing online social networks. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 32–39, October 2005. doi:10.1109/INFVIS.2005.1532126. 62

[92] Jeffrey Heer. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, 116(6):1844–1850, 2019. 113

[93] Louis Henry and Michel Fleury. Des registres paroissiaux a l'histoire de la population: Manuel de dépouillement et d'exploitation de l'état civil ancien. *Population (French Edition)*, 11(1):142–144, 1956. doi:10.2307/1525715. 3

[94] Nathalie Henry, Jean-Daniel Fekete, and Michael J. McGuffin. NodeTrix: A Hybrid Visualization of Social Networks. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1302–1309, November 2007. doi:10.1109/TVCG.2007.70582. xi, 28, 29

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

[96] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, and Sabrina Kirrane et al. Knowledge graphs. *ACM Comput. Surv.*, 54(4), July 2021. doi:10.1145/3447772. 43

[97] Infovis SC policies FAQ. 102

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

[99] J. David Johnson. UCINET: A software tool for network analysis. *Communication Education*, 36(1):92–94, January 1987. doi:10.1080/03634528709378647. 8, 29

[100] Frédéric Kaplan. The Venice Time Machine. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, DocEng '15, page 73, New York, NY, USA, September 2015. Association for Computing Machinery. doi:10.1145/2682571.2797071. 6, 19

[101] Karine Karila-Cohen, Claire Lemercier, Isabelle Rosé, and Claire Zalc. Nouvelles cuisines de l'histoire quantitative. *Annales. Histoire, Sciences Sociales*, 73(4):773–783, December 2018. doi:10.1017/ahss.2019.90. 1, 35, 36, 41

[102] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual Analytics: Definition, Process, and Challenges. In Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North, editors, *Information Visualization: Human-Centered Issues and Perspectives*, Lecture Notes in Computer Science, pages 154–175. Springer, Berlin, Heidelberg, 2008. doi:10.1007/978-3-540-70956-5_7. xi, 6, 15

[103] Daniel A Keim. Visual Analytics. page 6. 29

[104] Florian Kerschbaumer, Linda Keyserlingk, Martin Stark, and Marten Düring. *The Power of Networks. Prospects of Historical Network Research*. January 2015. 1, 2, 11, 26, 109

[105] Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and cellular networks. *PLoS computational biology*, 5(5):e1000385, 2009. 80

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

[107] Elena V. Konstantinova and Vladimir A. Skorobogatov. Application of hypergraph theory in chemistry. *Discrete Mathematics*, 235(1-3):365–383, May 2001. doi:10.1016/S0012-365X(00)00290-9. 80

[108] C. Kosak, J. Marks, and S. Shieber. Automating the layout of network diagrams with specified visual organization. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(3):440–454, March 1994. doi:10.1109/21.278993. xi, 28

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

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

[111] Charles-Victor Langlois and Charles Seignobos. *Introduction aux études historiques*. ENS Éditions, February 2014. 1, 16

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

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

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

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

[116] Claire Lemercier. 12. Formal network methods in history: Why and how? In Georg Fertig, editor, *Social Networks, Political Institutions, and Rural Societies*, volume 11, pages 281–310. Brepols Publishers, Turnhout, January 2015. doi:10.1484/M.RURHE-EB.4.00198. 1, 7, 11, 23, 25, 26, 27, 30, 33, 41, 52, 107

[117] Claire Lemercier and Claire Zalc. *Quantitative Methods in the Humanities: An Introduction*. University of Virginia Press, March 2019. 2, 3, 7, 18, 19, 27, 34, 35, 36, 41, 53, 109

[118] Claire Lemercier and Claire Zalc. Back to the Sources: Practicing and Teaching Quantitative History in the 2020s. *Capitalism*, 2(2):473–508, 2021. doi:10.1353/cap.2021.0010. 7, 18, 33, 34, 35, 36, 41

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

[120] Carola Lipp. Kinship Networks, Local Government, and Elections in a Town in Southwest Germany, 1800-1850. *Journal of Family History*, 30(4):347–365, October 2005. doi:10.1177/0363199005278726. 25

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

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

[123] Gribaudi Maurizio. *Espaces, Temporalités, Stratifications :. Exercices Méthodologiques Sur Les Réseaux Sociaux*. Editions de l'Ecole des Hautes Etudes en Sciences Sociales, Paris, January 2000. 23

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

[125] Fintan McGee, Benjamin Renoust, Daniel Archambault, Mohammad Ghoniem, Andreas Kerren, and Bruno Pinaud et al. *Visual Analysis of Multilayer Networks*. Synthesis Lectures on Visualization. Morgan & Claypool Publishers, 2021. 43

[126] Michael J. McGuffin. Simple algorithms for network visualization: A tutorial. *Tsinghua Science and Technology*, 17(4):383–398, August 2012. doi:10.1109/TST.2012.6297585. 28

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

[128] R. Michalski, P. Kazienko, and D. Krol. Predicting Social Network Measures Using Machine Learning Approach. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1056–1059, Istanbul, August 2012. IEEE. doi:10.1109/ASONAM.2012.183. 111

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

[130] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. Lulu.com, 2019. 86

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

[132] J. L. Moreno. *Who Shall Survive?: A New Approach to the Problem of Human Interrelations*. Who Shall Survive?: A New Approach to the Problem of Human Interrelations. Nervous and Mental Disease Publishing Co, Washington, DC, US, 1934. doi:10.1037/10648-000. xi, 22, 27

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

[134] Zacarias Moutoukias. Buenos Aires, port between two oceans: Mobilities, networks, stratifications (2nd half of the 18th century). *E-SPANIA-REVUE ELECTRONIQUE D ETUDES HISPANIQUES MEDIEVALES*, 25, 2016. 37, 58

[135] Zacharias Moutoukias. Réseaux personnels et autorité coloniale : Les négociants de Buenos Aires au XVIIIe siècle. *Annales. Histoire, Sciences Sociales*, 47(4-5):889–915, October 1992. doi:10.3406/ahess.1992.279084. 25

[136] Andrej Mrvar and Vladimir Batagelj. Analysis and visualization of large networks with program package Pajek. *Complex Adaptive Systems Modeling*, 4(1), April 2016. doi:10.1186/s40294-016-0017-8. 29, 61

[137] Natural earth. 61

[138] Neo4j graph data platform. 53, 55, 71, 80

[139] Mark Newman. *Networks*. Oxford university press, 2018. 22

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

[141] Carolina Nobre, Marc Streit, and Alexander Lex. Juniper: A Tree+Table Approach to Multivariate Graph Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):544–554, January 2019. doi:10.1109/TVCG.2018.2865149. 28

[142] Gérard Noiriel. Naissance du métier d'historien. *Genèses. Sciences sociales et histoire*, 1(1):58–85, 1990. doi:10.3406/genes.1990.1014. 16

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

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

[145] John F. Padgett and Christopher K. Ansell. Robust Action and the Rise of the Medici, 1400-1434. *American Journal of Sociology*, 98(6):1259–1319, May 1993. doi:10.1086/230190. xi, 1, 2, 5, 25, 26, 109

[146] Pajek — Analysis and visualization of very large networks. 6, 8, 84

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

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

[149] Vanessa Peña-Araya, Tong Xue, Emmanuel Pietriga, Laurent Amsaleg, and Anastasia Bezerianos. HyperStorylines: Interactively untangling dynamic hypergraphs. *Information Visualization*, 21(1):38–62, January 2022. doi:10.1177/14738716211045007. 47

[150] Cindarella Sarah Maria Petz. *On Combining Network Research and Computational Methods on Historical Research Questions and Its Implications for the Digital Humanities*. PhD thesis, Technische Universität München, 2022. 2, 26, 36, 109

[151] James P. Philips and Nasseh Tabrizi. Historical Document Processing: Historical Document Processing: A Survey of Techniques, Tools, and Trends, September 2020. arXiv:2002.06300, doi:10.48550/arXiv.2002.06300. 111

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

[153] Alexis Pister, Paolo Buono, Jean-Daniel Fekete, Catherine Plaisant, and Paola Valdivia. Integrating Prior Knowledge in Mixed-Initiative Social Network Clustering. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1775–1785, February 2021. doi:10.1109/TVCG.2020.3030347. 9, 29

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

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

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

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

[158] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. 55, 84

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

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

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

[162] Pedro Ribeiro and Fernando Silva. Discovering Colored Network Motifs. In Pierluigi Contucci, Ronaldo Menezes, Andrea Omicini, and Julia Poncela-Casasnovas, editors, *Complex Networks V*, Studies in Computational Intelligence, pages 107–118, Cham, 2014. Springer International Publishing. doi:10.1007/978-3-319-05401-8_11. 54

[163] Christian Rollinger. Prolegomena. Problems and Perspectives of Historical Network Research and Ancient History. *Journal of Historical Network Research*, 4:1–35, May 2020. doi:10.25517/jhnr.v4i0.72. 7, 8, 11, 42, 107

[164] Giulio Rossetti and Rémy Cazabet. Community discovery in dynamic networks: A survey. *ACM computing surveys (CSUR)*, 51(2):1–37, 2018. 26, 87, 109

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

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

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

[168] C.J. Rueda and Catedral de Buenos Aires. *Matrimonios de La Catedral de Buenos Aires, 1747-1823*. Number v. 2 in Fuentes Históricas y Genealógicas Argentinas. Fuentes Históricas y Genealógicas Argentinas, 1989. 37

[169] Anni Sairio. Methodological and practical aspects of historical network analysis: A case study of the Bluestocking letters. In Arja Nurmi, Minna Nevala, and Minna Palander-Collin, editors, *Pragmatics & Beyond New Series*, volume 183, pages 107–135. John Benjamins Publishing Company, Amsterdam, 2009. doi:10.1075/pbns.183.08sai. 43

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

[171] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. Vega-lite: A grammar of interactive graphics. *IEEE Trans. Vis. Comput. Graphics*, 23(1):341–350, 2016. 15, 71

[172] Shrutika S. Sawant and Manoharan Prabukumar. A review on graph-based semi-supervised learning methods for hyperspectral image classification. *The Egyptian Journal of Remote Sensing and Space Science*, 2018. doi:10.1016/j.ejrs.2018.11.001. 87

[173] John Scott. Social Network Analysis. *Sociology*, 22(1):109–127, February 1988. doi:10.1177/0038038588022001007. 11, 21, 22, 29, 40

[174] Vanessa Serrano Molinero, Benjamin Bach, Catherine Plaisant, Nicole Dufournaud, and Jean-Daniel Fekete. Understanding the use of the vistorian: Complementing logs with context mini-questionnaires. In *Visualization for the Digital Humanities Workshop*, Phoenix, United States, October 2017. xi, 28, 29, 30, 39

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

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

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

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

[179] Ben Shneiderman. Inventing Discovery Tools: Combining Information Visualization with Data Mining. *Information Visualization*, 1(1):5–12, March 2002. doi:10.1057/palgrave.ivs.9500006. 4, 85

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

[181] Marc A. Smith, Ben Shneiderman, Natasa Milic-Frayling, Eduarda Mendes Rodrigues, Vladimir Barash, Cody Dunne, Tony Capone, Adam Perer, and Eric Gleave. Analyzing (social media) networks with NodeXL. In John M. Carroll, editor, *Proceedings of the Fourth International Conference on Communities and Technologies, C&T 2009, University Park, PA, USA, June 25-27, 2009*, pages 255–264. ACM, 2009. doi:10.1145/1556460.1556497. 6, 29, 53, 61

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

[183] John T. Stasko, Carsten Görg, and Zhicheng Liu. Jigsaw: Supporting investigative analysis through interactive visualization. *Inf. Vis.*, 7(2):118–132, 2008. doi:10.1057/palgrave.ivs.9500180. 31, 46, 47, 110

[184] Chris Stolte, Diane Tang, and Pat Hanrahan. Polaris: A System for Query, Analysis, and Visualization of Multidimensional Relational Databases. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, 8(1):14, 2002. 47

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

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

[187] Shazia Tabassum, Fabiola S. F. Pereira, Sofia Fernandes, and João Gama. Social network analysis: An overview. *WIREs Data Mining and Knowledge Discovery*, 8(5):e1256, 2018. doi:10.1002/widm.1256. 1, 5, 22

[188] Melissa Terras. Quantifying digital humanities. *UCL Centre for Digital Humanities*, 2011. 19

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

[190] Charles Tilly. Retrieving european lives. 1984. 2, 11, 15, 34, 52

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

[192] Natkamon Tovanich, Alexis Pister, Gaelle Richer, Paola Valdivia, Christophe Prieur, Jean-Daniel Fekete, and Petra Isenberg. VAST 2020 Contest Challenge: GraphMatchMaker: Visual Analytics for Graph Comparison and Matching. *IEEE Computer Graphics and Applications*, pages 1–1, 2021. doi:10.1109/MCG.2021.3091955. 54, 55

[193] Francesca Trivellato. Is There a Future for Italian Microhistory in the Age of Global History? *California Italian Studies*, 2(1), 2011. doi:10.5070/C321009025. 18, 36

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

[195] John W. Tukey. *Exploratory Data Analysis*. Pearson, Reading, Mass, 1er édition edition, January 1977. 5, 15

[196] Paola Valdivia, Paolo Buono, Catherine Plaisant, Nicole Dufournaud, and Jean-Daniel Fekete. Analyzing Dynamic Hypergraphs with Parallel Aggregated Ordered Hypergraph Visualization. *IEEE Trans. Visual. Comput. Graphics*, 27(1):1–13, January 2021. doi:10.1109/TVCG.2019.2933196. 28, 47, 82, 89, 95, 111

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

[198] Ingeborg van Vugt. Using multi-layered networks to disclose books in the republic of letters. *Journal of Historical Network Research*, 1(1):25–51, October 2017. 43

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

[200] VisMaster: Visual analytics — Mastering the information age. 103

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

[202] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, November 1994. 5, 23, 29

[203] Charles Wetherell. Historical Social Network Analysis. *Int Rev of Soc His*, 43(S6):125–144, December 1998. doi:10.1017/S0020859000115123. 1, 3, 11, 24, 25, 34, 40, 43, 52, 109

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

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

[206] Hadley Wickham and Maintainer Hadley Wickham. The ggplot package. *Google Scholar*, 2007. 15

[207] Leland Wilkinson. *The Grammar Of Graphics*. Springer, New York, 1993. 14

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

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

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

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

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