

Thesis

Alexis Pister

June 27, 2022

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 5 |
| 1.1 | Social History and Historical Social Network Analysis | 6 |
| 1.2 | Network models and representations | 7 |
| 1.3 | Usability Issues | 8 |
| 1.4 | Contribution and research statement | 9 |
| 2 | Historical Network Analysis and Visualization | 11 |
| 2.1 | Social Network Analysis | 11 |
| 2.1.1 | Sociometry to SNA | 11 |
| 2.1.2 | Structuralism and Ego Studies | 12 |
| 2.1.3 | Methods dans tools | 13 |
| 2.2 | Historical Network Research | 13 |
| 2.2.1 | Social History | 13 |
| 2.2.2 | Historical Social Network Analysis | 14 |
| 2.2.3 | Network Modeling | 15 |
| 2.3 | Social Network Visualization | 15 |

Chapter 1

Introduction

Social scientists such as historians and sociologists want to make sense of the structure and dynamics of the social relationships between people of a given place and time. Social Network Analysis (SNA) and its history equivalent Historical Social Network Analysis (HSNA) is one of the main paradigm to achieve this task. It consists in constructing a network representing the social ties between the persons of interest, and studying this network to make sociological conclusions. Usually, persons are represented as nodes in the network, while the links model social relationships, such as friendships or family links. For this, social scientists try to exhaustively list all the persons in a restricted time and place with all their social ties, and create a network from it. The resulting network is considered to be a good model of the social reality, thus allowing us to study the structure and dynamics of the social fabric of a period, by studying the network in itself. In parallel, a lot of work has been done in network visualization and specifically Social Network Visualization (SNV) to make useful representations of social networks, and visual analytics tools allowing an effective exploration and analysis of this type of data. However, sociology and history data can be quite complex, and simple networks model are often a simplification of the real world social phenomena. Several network models ranging from simple to complex ones have been introduced, with associated visual representations and tools. But there is no consensus of which is better, specifically for networks constructed from historical sources, and the majority of research is still done using very simple models, with a classical node-link representation. Furthermore, current SNA tools such as Gephi or Pajek do not provide much guidance to the social scientists for their analysis, which often require a good computer science and statistics background. This thesis aim is to tackle those two

problems, by first defining a network model which models well most of the historical sources we encountered, and proposing visual representations to explore them. Then, we propose visual analytics tools and methods specifically designed for social scientists to explore their data, with the aim of proposing the right balance between algorithmic power, interpretation of the analysis and the decision.

1.1 Social History and Historical Social Network Analysis

History is the science of retrieving and characterizing facts about the past, in all their complexity. Traditional history methodology consists in finding and expliciting specific events—such as wars or diplomatic tensions—and eliciting their causes and consequences, and narrating the lives of historic figures, such as reigners or artists. But in the first half of the 20th century, a new history approach and methodology emerged called social history. This branch of history studies the socio-economic dynamics between the different groups of a society, instead of focusing on the affairs of a state. More recently, with the development of network science and computer science, sociologists started to study social phenomena and relationships from a network perspective. A network is an abstraction used to modelize phenomena based on relationships between entities, made of nodes and links. By modeling the social ties of a group of interest, such as preschooler [REF], or a karate club [REF] with the use of a network, sociologists can leverage quantitative measures from the network to make sociological conclusions. This network analysis approach grew in popularity in recent years, and has started to be used and formalized by historians, under the term of Historical Network Research (HNR). Similarly to sociologists, historians can build a network modeling the social relationships of actors of the past, restricted in a specific period and area they are studying. If sociologists can use surveys, experiments or nowadays the internet to extract social relationships and construct a social network, historians are restricted by the written sources they can find. Their main source of work to extract social relationships in a rigorous way are historical documents which correspond to traces of specific events linking people together. These can be marriage acts, birth certificates, or census for family and close personal relationships, or migration acts and working contracts for other types of social ties. After having a selected corpus, they have to annotate manually each document to extract the persons mentioned in it along the relationships between them, to finally construct a network from

this data. This is a long and tedious process which can result in small to large networks that they want to analyze to make conclusions on the social dynamics of a population of interest. For this complicated task, historians follow what is called a Social Network Analysis (SNA), or more precisely a HHistorical Social Network Analysis (HSNA) which consists in characterizing the structure of the network with measures such as the centrality or the density of some parts of the network to then make conclusions on how people were interacting in the period of interest. To help their analysis, and generate new hypotheses, they usually rely on Visual Analytics tools to represent and explore their network. The elaboration of visual tools to represent and explore social networks is called Social Network Visualization. Sociologists and historians started to use static representation of networks, using node-links diagrams to have a visual understanding of their data, and to report their findings in publications. With the development of visualization, more complex representations and visual analytics tools emerged, which allow more complex representation and exploration capabilities, with the help of interactions and navigations features. Social networks visual systems such as Gephi or Pajek are now widely used in HNR and SNA by social scientists. Representing their network data and being able to interact with it allows them to rapidly have an overview of it, confirm hypotheses they have and arrange new ones by exploring the network.

However, most used social networks visual analysis tools still have several issues that we tackle in this thesis : the visual representations still widely used are pretty simple, and are often not a good fit to represent and explore complex multivariate historical dataset, and current visual analytics tools often do not provide enough power and guidance to the end users to manipulate their data, which can result in frustration.

1.2 Network models and representations

Person-to-person simple node-link diagram is still the most widely used network representation is SNA, and most SVA tools only include this type of representation. This visualization shows the persons as nodes, and social ties as links and displays them in a way to minimize the number of crossings to increase the readability. However, historians very often have access to richer and more diverse information through the historical documents they study. The documents can refer to coexistent complex social relationships which link several people together with different roles. These cannot be modeled with simple person-to-person links, without losing some informa-

tion on the social implication of these relationships. Moreover, documents often give access to other information related to the event they refer to, such as the time, the location or the roles of the different persons mentioned. For example, marriage acts often indicate the date and the place of the event, and mention persons under different roles : the spouses, the witness, the parents, the priest etc. Additional information related to persons can also be mentioned, such as their age, origin or profession. It is clear that simply using a person network model won't encapsulate the whole complexity of the data and will simplify the social relationship. This is a common issue in SNA and HSNA [REF Lemercier] and more complex network models are needed. However, complexity along with visual analysis tools to explore them.

1.3 Usability Issues

One of the aims of Visual Analytics is to provide automatic or semi-automatic processing and analysis tools with data mining and machine learning algorithms, to help end users make sense of their data and find interesting patterns and relationships. However, current social network visual analytics systems are still very algorithm oriented, and do not provide many controls to historians and sociologists who usually feel off the analysis loop when the system provides automatic and algorithmic results. One of the reason is because automatic results can be hard to interpret, especially in a discipline such as History or Sociology, where users often have little knowledge on computer science. One example is the automatic detection of community structures using network clustering algorithms. Social networks are known to have a community-like structure, meaning that the probability of a link existing between two random person nodes is not uniform, and that people tend to agglomerate in groups, who have more social ties between them than with other persons in the network. There are a lot of existing clustering algorithms which aim to automatically find these groups, by optimizing measures such as the modularity or using propagation models. However, clustering is an ill-defined problem, and several good partitions may coexist for the same network, and which can have several interpretations in a SNA. Most SNA/SVA tools such as Gephi or NodeXL provide several well known clustering algorithms such as Girvan-Newman, Louvain or Clauset-Newman-Moore, but do not provide much guidance on how to use them and interpret their results. Social Scientists often try several ones in the list of algorithms proposed until finding a convenient result, in the eyes of the anal-

ysis they want to follow. This leads to a non satisfactory analysis process, as historians are out of the loop and have few decisions on the results. This usability issue is the same for automatic processes with no universal ground truth.

1.4 Contribution and research statement

This thesis is centered around two research questions: first, the proposition of an efficient network model to represent historical sources as a network, with associated visualizations to show and explore this type of network. Secondly, elaborating visual analytics tools to explore this type of data with the right balance of algorithmic power, simplicity and interpretability for the social scientists, who need to be in control of the analysis. We first [tell plan]

Chapter 2

Historical Network Analysis and Visualization

2.1 Social Network Analysis

2.1.1 Sociometry to SNA

Sociology always had the goal of studying social relationships between individuals, and finding recurrent patterns and structures allowing to describe and explain the behaviour of people and groups. Traditional methods and paradigm saw and explained social phenomena through the lens of social groups and categories, such as age, social status, profession and sex. For example, the social position of people living in a small city could be explained well by their age, demographics and social status which are traditional social categories. However, some critics emerge that the division in social categories is often partially biased and come from predefined divisions which are not always grounded in reality. Sociometry which is considered as one of basis of SNA had the goal of redefining social categories through the lens of real social interactions and ties between persons. It is in the 1930s that Moreno started to develop this discipline by trying to depict real social interactions as a way to understand how groups and organization were functioning. For this, he developed sociograms as a way to visually show friendships between people with the help of circles representing persons and lines modeling friendships. This way, he could rapidly see the different main actors and hubs of interaction inside the social fabric represented visually. Sociometry tremendously helped disseminate the metaphor of network to model and understand social phenomena, especially with the help of car-

tograms. It was until the 1960s that sociologists took these concepts further and formalized SNA using graphs and mathematical methods. It followed the emergence of Graph Theory studies in the 1950 by Mathematician such as Erdős [...]. It did not take long until sociologists used these concepts to model social ties and relationships into graphs. Sociologists already have structural theories of social phenomena, and they rapidly saw the potential of graphs to model and analyze those in a mathematical way. Several sociologists started to codify those concepts to use them in a sociology setting such as Coleman (1964). They started to model social ties between agents as graphs $G = (V, E)$ with V a set of vertices representing agents such as persons and organizations, and $E \subseteq V^2$ a set of edges modeling the social ties between pairs of agents. Once social phenomena were modeled as networks, a variety of methods and measures coming from graph theory such as the centrality or the diameter were at their disposal. Sociologists started to make links between these measures and sociological facts. It was then possible to make sociological conclusions from the direct observations of social ties modeled as networks.

2.1.2 Structuralism and Ego Studies

After SNA started to be formalized, lots of sociological studies have been done using those concepts. However, there was not yet strong protocols and methods to follow, and networks are an abstraction that can model different things in different ways. When looking retrospectively, we can see that two schools of thought emerged with different objectives and methods: the structuralists and the school of Manchester.

The Structural Analysis of Social Networks refers to the Structuralists in Sociology. They are interested in the properties and structure of the network, and make parallel between them and how persons were interacting in real life. They think the position of persons in the network and the relational patterns they are part of reflect well the social activities and behavior in real life. Accordingly, sociologists in this school usually study organization and specific groups, and want to explain their behavior and interaction through the internal shape and structures of resulting networks. They thus try to construct network which exhaustively model all the interactions between the actors constituting the groups.

In contrast, the school of Manchester tries to explain specific persons' behavior and social interactions, through their direct interactions and without necessarily studying a global network structure. This school of thought is related to the concept of ego networks. Ego networks consist in all the di-

rect relation of one node—in this case a person—with the relation between persons of this small network. They usually want to model the different types of relationships of a person, like their family, work and friends ties and study them through time. They make a direct parallel between these direct social ties and the status, condition and life of persons, and usually compare several ego networks to make conclusions about the correlations between the two.

These two ways of seeing SNA are often not exclusive and current studies usually involve concepts and methods from these two schools.

2.1.3 Methods and tools

Graph theorists and network scientists developed a myriad of measures and algorithms that sociologists appropriated themselves to describe and characterize social phenomena. When constructing networks, the first thing sociologists did was often to identify the main actors of the network. Computing the degree (number of connection) distribution is the main straightforward way of doing it, but other measures like the centrality have been developed too. Centrality aimed at characterizing the most important actors of the network, and several centrality measures have been proposed, based on different criteria. Centrality can highlight actors with the highest number of connections, or those bridging different groups with low interactions. The concepts of dyads and triads counting which are simple structural elements give insight on how people are socializing and reflects on Simmel formal sociology, where he already referred as dyads and triads as primal form of sociability. More recently, the concept of graphlet extended this concept to every pattern of N-entities. Graphlet analysis aims at enumerating every small structure of N nodes composing a network, to understand how people interact at a low-level.

Groups/Clustering

2.2 Historical Network Research

2.2.1 Social History

Historians try to understand an epoch using textual sources from the past, and trying to extract useful information from them. Social history, which is a branch of history, focus on understanding how societies were organised and how people were living together at a particular time and place. Charles Tilly argued that the task of social history lays in "(i) documenting large

structural changes, (2) reconstructing the experiences of ordinary people in the course of those changes, and (3) connecting the two”. For the latter, historians can leverage personal written sources—such as letters, journals, books, and newspapers—to have the internal point of view of persons living in this society and descriptions of lives of precise individuals. For the former, historians usually need to study more structured documents which contain information which can be extracted in a predefined and exhaustive way. These documents can for example be census, migration acts or marriage acts. By studying these documents and by systematically extracting the information of these documents, historians can make global and quantitative conclusions on certain social and behavioural aspects of societies of interest. For example .. [EXAMPLE CHANGEMENT METIERS XXth century]

2.2.2 Historical Social Network Analysis

History started to adopt some of the methods and vocabulary of Network research in the 1980s, several years after other fields such as Sociology or Anthropology (TO CHECK). Before that, historians were already describing relational structures when studying families and organization. It was often a part of discussion and a conclusion of several studies. Network research was a way to put these relational structures as an object of study in itself, and allowed to study them in a more systematic and quantitative way. Instead of only looking at classes and groups, historians thus started to look at relational links between individuals, such as family, friendships or business ties. They already had techniques and tools to annotate and extract quantitative information from textual sources that they adapted to extract and study social ties. We therefore saw the emergence of HNR studies, where historians followed HSNA studies on networks constructed from the mention of social ties of their textual sources. It allowed them to make observations on previous objects of study like families or organization that it was not possible to see without taking into account the relational aspects of these phenomena. However, constructing a network from historical sources, which can differ in their structure is not a trivial task. The most straightforward approach, based on the most well known social network analysis, consists in constructing social network based on simple graph $G = (V, E)$ with V a set of vertices representing the persons of interest, and $E \subseteq V^2$ a set of edges modeling the social ties between pairs of persons. This allows to have a simple network to visualize and analyze, but does not always reflect the social complexity of the real relationships. More complex networks models have been proposed in SNA to be able to model more complex social

relationships.

2.2.3 Network Modeling

2.3 Social Network Visualization

Bibliography