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

Analyse Visuelle pour l'Analyse de Réseaux Sociaux Historiques

Visual Analytics for Historical Network Research

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

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

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

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

Alexis PISTER

Composition du jury

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

THESE DE DOCTORAT

NNT : 2020UPASA001

université
PARIS-SACLAY

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

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

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

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

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

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

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

Table des matières

1 - Historical Network Analysis and Visualization

Social historians rely on textual historical documents to draw socio-economic conclusions about the past. They read and analyze all the documents they can find from a period and subject of interest, and make their conclusions after analyzing them and cross-referencing the information they found. During this process, they can use several methods developed in History to extract and analyze the information contained in the documents in a scientific way, such as qualitative analysis, quantitative methods or HSNA. HSNA is a method coming from Sociology consisting in modeling the relational information mentioned in the documents—such as familiy, business or friendship ties—in a network, to be able to characterize and explain social behaviours through the description of the structure of the network. Historians following HSNA processes have inspired their workflow from Social Network Analysis (SNA), which is a well-known method in sociology and where a lot of methods and protocols had already been proposed when historians started to use similar approaches. Historians appropriated themselves this method and adjusted it to historical workflow which can vary from sociology as historians are limited in the documents they have and by their structure. They first have to annotate the documents to extract useful information, to then model it into an analyzable network. The annotation and modeling process is thus particularly complicated and specific to HSNA. Historians usually use social network visualization tools to confirm or generate new hypothesis once they successfully constructed their network. As the models used by historians are more and more complicated, new visualization systems are needed, first to analyze their networks, but also to help them in their HSNA process, from the acquisition of relevant documents to the final analysis and visualization steps.

1.1 . Social Network Analysis

The concept of SNA emerged in sociology in response to traditional methods using pre-defined taxonomies and social categories to understand and explain sociological behaviours and phenomena, which could introduce bias. By modeling real observed social relationships and interactions with networks and by using mathematical and statistical methods to study those, sociologists have been able to explain sociological phenomena and describe sociological interactions through their direct observation and manipulation. SNA is now a well praised methodology in sociology, which have also been appropriated by historians to study relational aspects of societies and institutions of the past.

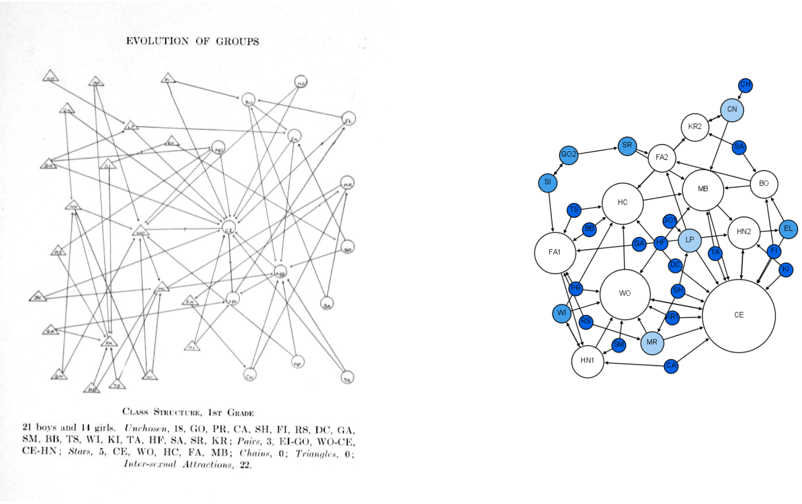
1.1.1 . Sociometry to SNA

One of Sociology's main goal is to study social relationships between individuals and finding recurrent patterns and structures allowing to explain the behaviours of people and groups. Traditional methods try to explain social phenomena using classical social classifications such as the 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 criticism emerged that this type of division is often partially biased and come from predefined categories which are not always grounded in reality. Sociometry is considered as one of the basis of SNA and had the goal of redefining social categories through the lens of real social interactions and ties between persons, that sociologists wanted to observe in real conditions. It is in the 1930s that Moreno started to develop this new method by trying to depict real social interactions as a way to understand how groups and organization were functioning [**?**]. He elaborated sociograms as a way to visually show friendships between people with the help of circles representing persons and lines modeling friendships. Sociometry tremendously helped disseminate the metaphor of networks to model and understand social structures and phenomena. It was during the 1960s that sociologists and anthropologists took these concepts further and formalized SNA using graphs and mathematical methods, following the emergence of Graph Theory studies in the 1950 by Mathematicians such as Erdos [**?**]. Sociologists already had structural theories of social phenomena, and they rapidly saw the potential of graphs to model social relationships between actors, representing the persons as nodes and relationships as links. Graph theory brought a panoply of concepts and methods to study and describe networks, that sociologists such as Coleman started to codify to use them in a sociology setting [**?**]. Using mathematical and network methods, it was possible to formally describe social relationships to make sociological conclusions grounded in real observations modeled as networks.

1.1.2 . Structuralism and Ego Studies

Lots of sociological studies used SNA concepts after it has been formalized. 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 thoughts emerged with different objectives and methods : the structuralists and the school of Manchester [**?**, **?**, **?**].

The structuralists are interested in observing the relational structures and patterns forming a network, to make parallels between them and the social behaviours of actors in real life [**?**]. They think the positions of the persons in the network and their relational patterns they are part of reflects well the social activities and behavior in real life. Studying those would thus allow them to make interesting sociological conclusions. Accordingly, sociologists in this school usually study organizations and specific groups—such as institutions and business companies—and want to explain their functioning through the description of the internal shapes and structures of the resulting networks. Thus, they try to construct networks which



EVOLUTION OF GROUPS

CLASS STRUCTURE, 1ST GRADE

21 boys and 14 girls. *Unchosen*, 18, GO, PR, CA, SH, FI, RS, DC, GA, SM, BB, TS, WI, KI, TA, HF, SA, SR, KR; *Pairs*, 3, EI-GO, WO-CE, CE-HN; *Stars*, 5, CE, WO, HC, FA, MB; *Chains*, 0; *Triangles*, 0; *Inter-sexual Attractions*, 22.

Figure 1.1 – Moreno original sociogram of a class of first grades from [**?**] (left). The diagram shows 21 boys (triangles) and 14 girls (circles). The same sociogram plot using modern practices generated from Gephi from [**?**]. The color encode the number of connections incoming.

exhaustively model all the interactions between the actors constituting the groups, as missing links would misrepresent the reality of interactions.

In contrast, the school of Manchester constituted by anthropologists focus on studying specific individuals and all their interactions in the different facets of their lives and in time. They typically want to explain certain behaviours and social characteristics of individuals by their relationships and interactions in all their complexity, and highlight the influence of some social aspects of one's life on other aspects. One famous example is Mayer's study on austral africa rural migrants going in cities [**?**] where he showed that the integration of urban mores and customs were directly correlated to the persons relationships networks in the city. Xhosa peoples still interacting with rural people of their village in the city were less changing their customs. This school of thought typically rely on the concept of ego network and more recently dynamic and multiplex networks. Ego networks are networks modeling all the direct relations of one central node—in this case a person—including the relations existing between the persons of this small network. They typically try to model the different types of relationships of a person, like their family, work and friendship ties and study them through time. By studying the ego network structure of someone, sociologists of this school try to leverage explanations on other social aspects of the persons like their social status, job and gender. It is also common to compare several ego networks to make correlations between the social relationships of individuals and other interesting social categories.

These two methodologies of SNA are often not exclusives and current studies usually involve concepts and methods from both.

1.1.3 . Methods dans 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, and explain why these actors were the most central, for example by linking it to their profession or social status. Computing the degree—which is the number of connections for a node—distribution is the main straightforward way of doing it, but other more complex measures like the centrality have also been developed. Lots of types of centrality have been proposed, based on different criteria, as there are several ways of defining the more <emph>important</emph> actors. Some centrality measures highlight actors with the highest number of connections while others highlight people bridging different groups with low interactions. More generally sociologists aimed at identifying recurring patterns of sociability between actors. The concepts of dyads and triads counting which are basic structural patterns of 2 and 3 nodes give insights on low level relationships between people. This reflects on Simmel formal sociology, where he already referred to dyads and triads as primal form of sociability [**?**]. More recently, graphlet analysis 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. Graphlets counting shows that graphlets are not found in an uniform distribution in social networks, thus revealing that these networks do not follow a random distribution. This is a fact well known in SNA. Precisely, entities in real world networks tend to agglomerate into groups (also called clusters) where entities in the same groups interact more between them than with entities from other groups. In a sociology perspective, it means that people tend to interact and socialize in groups, and interact more rarely with other people from outside groups. These groups are often referred as *communities*, and a lot of algorithms have been proposed to find these automatically.

1.2 . Historical Network Research

If Sociology and Anthropology started to use network concepts and methods rapidly in the 1950s, it was not until the 1980s that historians started to use this type of methodology. Yet, historians started to use quantitative methods from the 1930s, with the rise of social history, by extracting information from historical textual documents and studying them with statistical methods in the 1960s. When seeing the potential of SNA concepts for historical purposes, historians started to extract the relational information contained in documents to study historical social phenomena using the power of networks and methods already developed in SNA.

1.2.1 . Quantitative History

1.2.2 . Historical Social Network Analsyis

History started to use concepts and methods from SNA in the 1980s [**?**] in order to criticize quantitative history concepts and results, and to develop historical approaches—like *Microstoria* [**?**]—that focus on the study of individuals and groups through the lens of their interactions and relationships directly extracted from historical documents. Beforehand, historians were already describing and studying relational structures such as families and organizations with qualitative methods or with classical taxonomies, without studying in depth the relational aspect of these entities. Network research allowed to model those relational entities more thoroughly using networks concepts, thus allowing to make new observations that it was not possible to see without taking into account the relational aspects of these entities. example Lemercier Observing and describing the structure of the resulting networks allowed historians to make conclusions on sociological aspects of the past, similarly to SNA. Since then, HNR has been applied by sociologists and historians to study multiple kinds of relationships, like kinship and political mobilization [**?**], administrative and economic patronage [**?**], etc. If these approaches fall under similar critics of quantitative history [**?**] as for example the leading of trivial conclusions, it still led to classical works and interesting discoveries. One famous example is the study of the rise of the Medici family in Florence in the 15th century by Padgett [**?**], where he explained the rise of power of this family by their central position in the trading, marriage and banking networks of the powerful families of Florence. Figure 1.2 shows the different networks of Florence families where we can see the central position of the Medici. lots of historians are using and continuously improving the HNR method which can be very effective to study relational historical phenomena [**?**]. Moreover, historians rarely rely on a single approach when studying an era or phenomenon, they mix methods and tools from several domains of social and formal sciences with their own practices [**?**, **?**].

1.2.3 . Network Modeling

Constructing a network from historical documents, which can vary a lot in their formats and structures is not a trivial task. The most straightforward and well-known approach consists in constructing a social network based on a simple graph $G = (V, E)$ with V a set a vertices representing the actors of interest (very often individuals mentioned in the documents), and $E \subseteq V^2$ a set of edges modeling the social ties between pairs of actors. This allows to have a simple network to visualize and analyze, but it does not always reflect the sociological complexity of information contained in the documents. HNR network models have evolved over time to better take into account concrete properties of social networks, such as types of actors using labeled networks, the importance of actors or relations with weighted networks, mixed relationships with multiplex networks, dynamics of relations with dynamic networks. Bipartite networks have been proposed to

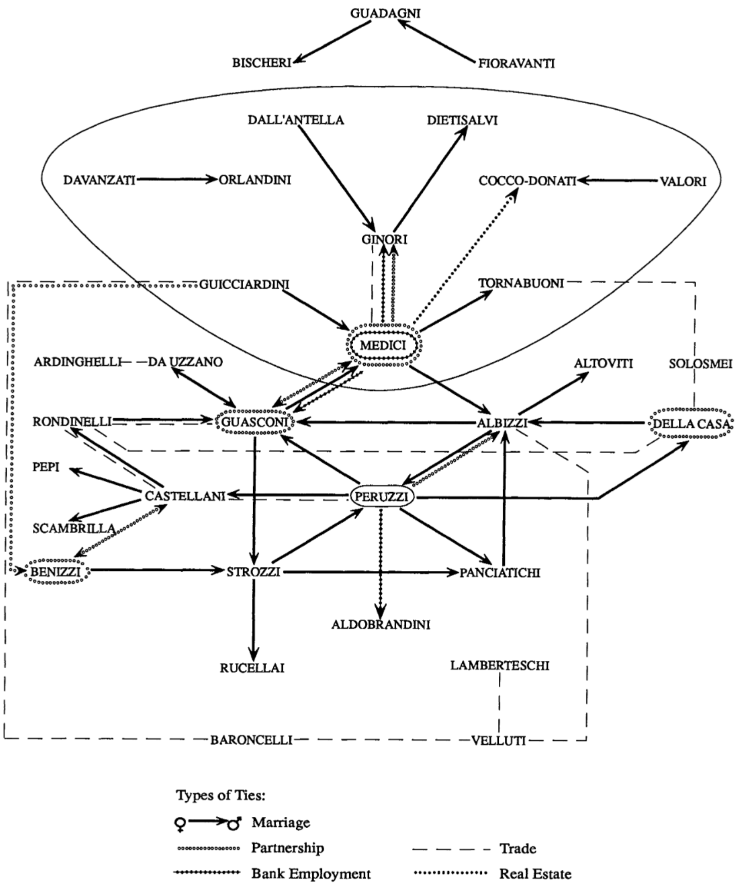


Figure 1.2 – Marriage, partnership. trading, banking and real estate networks of the powerful families of Florence from [**?**]. We can see the central position of the Medici Family

model relations between two types of entities, such as organization and employees where the relations link employees to organizations but not employees to employees or organizations to organizations. Many social situations or documents can be modeled in these terms (affiliation lists or co-authoring). Multivariate networks, i.e., graphs, where vertices and edges can be assigned multiple "properties" or "attributes", are less used in SNA. These attributes are often considered secondary, the emphasis of SNA being on the topology, its features, measures, and evolution.

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

1.3 . Social Network Visualization

Practitioners of SNA and HNR have always depicted visually their networks for validation and communication purposes, mostly using node-link diagrams. With the increase of average network size and the diversity of network models, new visualization techniques have been proposed to represent the diversity of studied networks. Moreover, more and more social scientists are now following exploratory approaches using Visual Analytics (VA) tools, to describe more in depth their data and generate new interesting hypothesis, using interaction and exploration capabilities.

1.3.1 . Visualization

Data Visualization consists in graphically displaying data in the purpose of enhancing human cognition capabilities to understand and communicate ideas and phenomena. History is filled with classical examples of visual data displays which helped understand real phenomena, such as Minard's map of Napoleon march in Russia [**?**], or Snow's dot map of cholera cases in London which showed the proximity between street pumps and cholera infections [**?**]. If several examples of data visualization can be found thorough history, it mainly developed as a scientific field in the 1960s with Tukey's work on data analysis and visualization [**?**] and Bertin publication of Semiology of graphics [**?**]. In this foundational work, Bertin described and organized the different visual elements usable in graphical information displays, and linked them to data features and relations types. Friendly says that "To some, this appeared to do for graphics what Mendeleev had done for the organization of the chemical elements" [**?**]. The development of computer science and the rise of hardware capabilities during the same time created a big need for data visualiza-

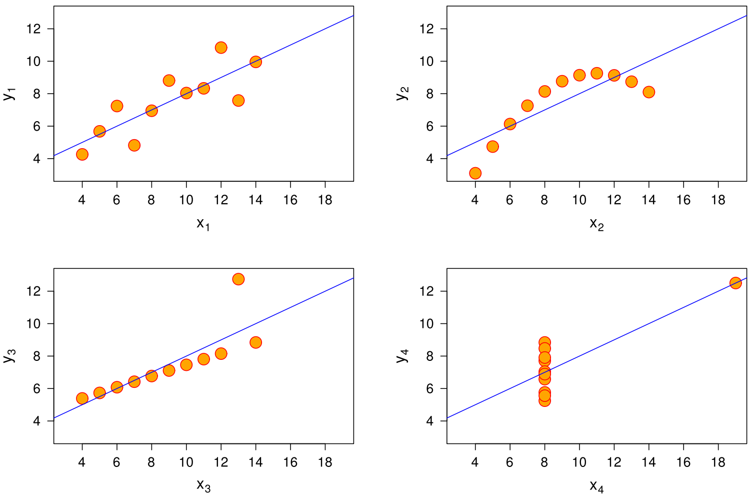


Figure 1.3 – Anscombe quartet. The four datasets have the same descriptive statistics (average, variance, correlation coefficient) but very different structures. Image from [**?**].

tion. The amount of data stored increase exponentially and descriptive statistics were not enough to understand the underlying structure of the amount and diversity of produced data. Visualization, leveraging the human visual system, allows to rapidly see the hidden structure of a dataset and detect interesting and unexpected patterns very often unseen with classical statistical methods. One classical illustration of this is Anscombe's quartet [**?**] which consists in four datasets of points in $\mathbb{R}^2$ with the same statistical measures (mean, variance, correlation coefficient etc.) but with very different structures, that plotting the data show immediately. The four datasets are illustrated in Figure 1.3.

Lots of visualization techniques emerged to make sense of the diversity of data produced, such as relational, temporal, spatial or network data. Subfields of Visualization emerged : **Scientific visualization** focus on visualizing continuous real data such as weather, spatial, and physics data, sometimes produced with simulations whereas **Information Visualization** is centered around the visualization of (multidimensional) discrete data points, often in an abstract way. **Visual Analytics** emerged later from Information Visualization by mixing data mining and more complex analysis process with traditional information visualization problematics. Historical Social Network visualization is closely related to Information Visualization and Visual Analytics, and good visualization systems for HNR use concepts and methodologies from those two fields.

1.3.2 . Social Network Visualization

Sociologists rapidly saw the potential of graphically showing relationships between individuals, to better comprehend the underlying social structure and communicate their findings. Moreno elaborated sociograms to visually show friendships among schoolchildren with circles and lines to respectively show children and friendships ties [**?**]. This type of representation—commonly called node-link diagram—is the most widely used in social sciences, as it is rapidly understandable and effective for small to medium-sized networks which is usually the norm in social sciences. The most used social network visual analytics software such as Gephi [**?**] and Pajek [**?**] are based on this type of representation and allow a fully integrated exploration and analysis with the help of various algorithms. Finding an optimal placement for the nodes is however not that simple as several metrics can be optimized depending on the desired drawing, such as number of edge crossings, the variance of edge length, orthogonality of edges etc [**?**, **?**]. Figure 1.4 shows some of these metrics, synthesized by Kosara and al. [**?**]. In Figure 1.1 we can see the difference in readability between the original manual layout (left) and an automatic one (right). Automatic layouts which aim at optimizing readability metrics give clearer diagrams. The number of edge crossings is often considered as the most important measure, but finding a drawing with the optimal number of crossing is a NP-Hard problem, meaning that heuristics are needed for most real world use cases. Lots of algorithms have been designed such as force-directed ones, modeling the nodes as particles which repulse each other and are attracted together when connected with a link which can be seen as strings. Other visual techniques have been proposed to represent networks such as matrices, circular layouts and arcs, but are less used in social sciences [**?**]. Still, Matrices have been shown to be better than node-link diagram for a lot of tasks such as finding cluster related patterns, especially for medium to large networks [**?**].

As social scientists started to use more complex network models such as bipartite or temporal networks, more sophisticated representations are needed. The visualization community developed new representations to visualize other network types such as dynamic hypergraphs with PAOHVis [**?**], clustered graphs with NodeTrix [**?**] (illustrated in Figure 1.5), geolocated social networks with the Vistorian [**?**], and multivariate networks with Juniper [**?**]. However, these new networks representations take time to be adopted by social scientists who rarely use those.

1.3.3 . Social Network Visual Analytics

read notes Social network visualization has mostly been used for confirmatory and communication purposes from its beginning. Social scientists often had hypothesis that they could rapidly verify by plotting the data. The same plots were often used for communication purposes, for example in a scientific paper or presentation. However, visualization can also be used for exploratory aims, to gain new insights on the data and potentially generate new hypothesis. This process has been characterized by Tukey in 1960 as *Exploratory Data Analysis (EDA)*. Exploration is mostly possible thanks to interaction, which allows to change the point of focus in the data to highlight interesting patterns, with the help of mechanisms like filtering, querying, sorting etc. As the average size of datasets keeps growing,

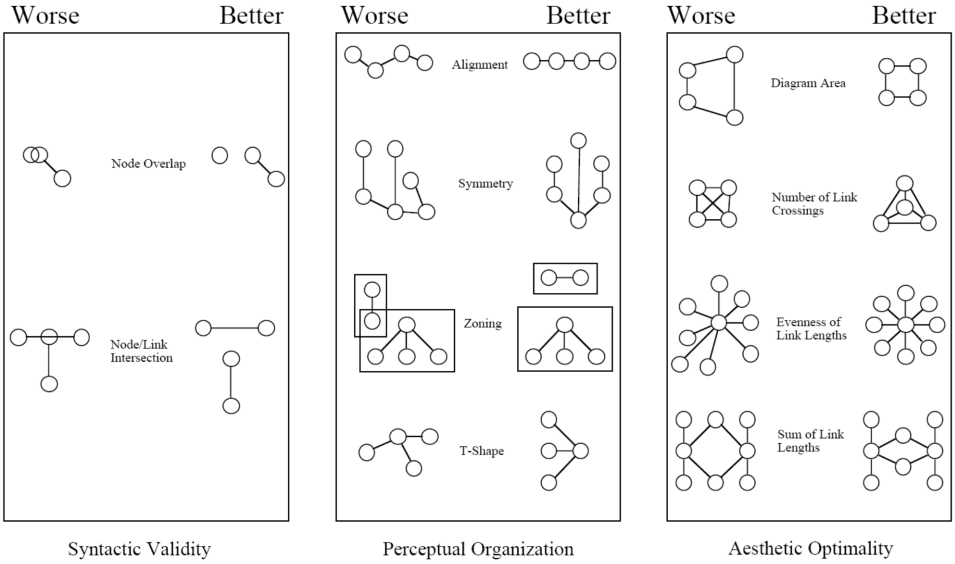


Figure 1.4 – Different criteria proposed to enhance node-link diagram readability. Image from [**?**]

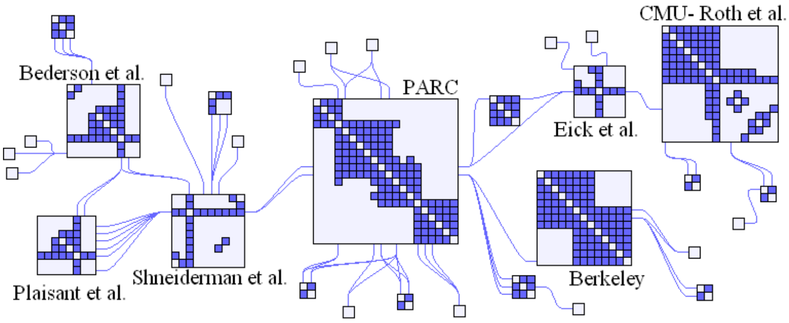


Figure 1.5 – NodeTrix system showing a scientific collaboration social network with clusters. Each cluster is represented as a matrix, Image from [**?**]

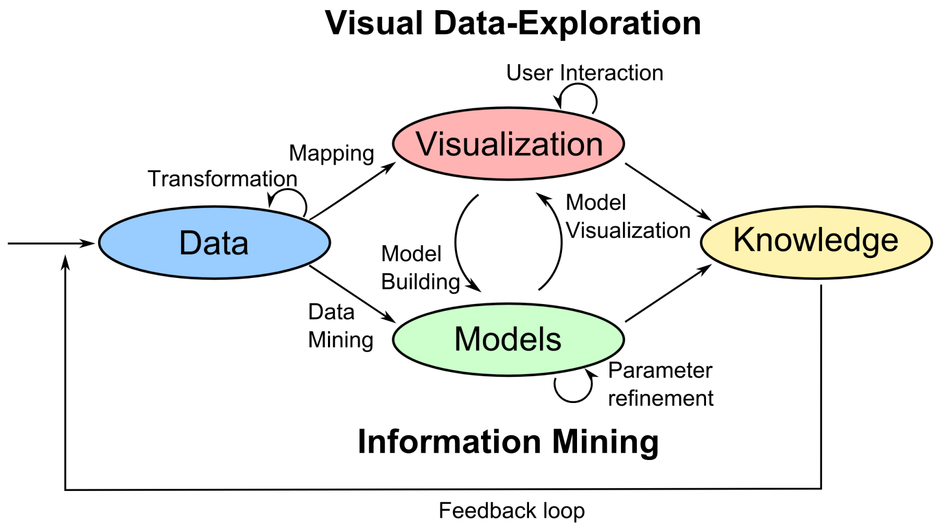


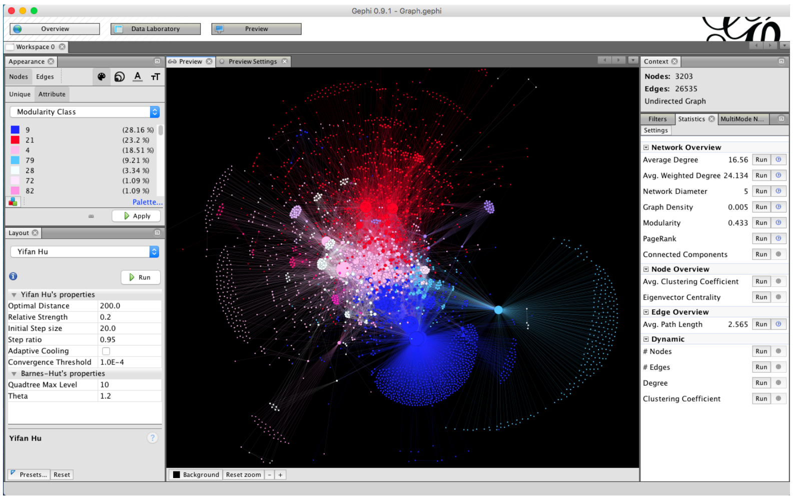
Figure 1.6 – Abstraction of the VA process. It is characterized by continuous interactions between the data, visualizations, models and knowledge. Image from [**?**]

exploratory tools are often needed to make sense of large datasets and generate interesting hypothesis.

Social scientists also often want to gain insight with the help of statistical and machine learning methods, that visualization only can not provide. More recent visual exploration interfaces incorporate automatic analytical tools along graphical displays, letting users apply data mining algorithms directly in the exploratory loop. This coupling of visualization and data mining has been defined as Visual Analytics (VA) and is still undergoing lots of research. Keim and al. define it as "a combination of automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data". Figure 1.6 shows an abstract representation of the VA process.

It is defined around the generation of knowledge using visualizations and models of the data, that the user generate and explore using interaction. Social scientists now frequently use VA systems to make sense of their data by using visualization, interaction, and data mining algorithms in their analysis loop to find interesting patterns and verify and create hypothesis. The most used social network VA tools are Gephi [**?**], Pajek [**?**] and NodeXl [**?**]. They all let users visualize their networks with a node-link diagram, and allow an interactive exploration of the data with operations like filtering. Users can also analyze their data using network measures computed directly in the interface, and apply data mining algorithms such as clustering which results are explorable visually.

Unfortunately, social scientists are often not trained in computer science and mathematical methods, and a lot of them have been frustrated by VA tools and



Modularity Class

🟦 9	(28.16 %)
🟥 21	(23.2 %)
4	(18.51 %)
79	(9.21 %)
28	(3.34 %)
72	(1.09 %)
82	(1.09 %)

Yifan Hu

Yifan Hu's properties

Optimal Distance	200.0
Relative Strength	0.2
Initial Step size	20.0
Step ratio	0.95
Adaptive Cooling	□
Convergence Threshold	1.0E-4

Barnes-Hut's properties

Quadtree Max Level	10
Theta	1.2

Yifan Hu

Nodes: 3203
Edges: 26535
Undirected Graph

Network Overview		
Average Degree	16.56	Run
Avg. Weighted Degree	24.134	Run
Network Diameter	5	Run
Graph Density	0.005	Run
Modularity	0.433	Run
PageRank		Run
Connected Components		Run
Node Overview		
Avg. Clustering Coefficient		Run
Eigenvector Centrality		Run
Edge Overview		
Avg. Path Length	2.565	Run
Dynamic		
# Nodes		Run
# Edges		Run
Degree		Run
Clustering Coefficient		Run

Figure 1.7 – Gephi [**?**] interface. The network is represented with a node-link diagram. Users can interact on the visualization and encode node and links visual attribute (color, size etc.) with network measures computed directly in the interface, such as the node degree, or clustering results.

by how it was guiding their analysis in predefined ways. For example, lots of social network VA interfaces propose clustering features, allowing users to find interesting groups with the help of automatic algorithms. However, social scientists often do not understand how the algorithms work and are not always satisfied with the results, as they can have knowledge from other sources not modeled inside the network. They usually end up trying several algorithms until they stumble upon a satisfactory enough solution. Cleaning and importing the data is also complicated, as the annotation and network modeling process are not straightforward and social scientists often encounter errors and inconsistencies in the data once they visualize it, that they would like to correct. Therefore, more work is still to be done on social network VA tools, to provide more guidance and power to social scientists while doing their analysis, and helping them to do easier back and forth between their analysis and the annotation, network modeling, and cleaning steps.