A Survey of Network Visualization in Humanities

Bachelor Thesis of

Rebecca Seelos

At the Department of Informatics
Institute of Theoretical Computer Science

Reviewers:	Prof. Dr. Dorothea Wagner
	Prof. Dr. Henning Meyerhenke
Advisors:	Dr. Tamara Mchelidze

Time Period: 1st December 2016 – 30th April 2017

Statement of Authorship

I hereby declare that this document has been composed by myself and describes my own work, unless otherwise acknowledged in the text.

Karlsruhe, 28th April 2017

Abstract

This thesis is a survey of network visualization techniques that are used in different fields of humanities. I will present rather recent examples that I found in various digital humanity and data visualization conferences of the last years. As the field of digital humanities is growing, this work focuses on how these techniques are used in particular in each humanity field.

Deutsche Zusammenfassung

Diese Arbeit soll eine Übersicht über Anwendungen von Techniken der Graphenvisualisierung in den Geisteswissenschaften geben. Dafür präsentiere ich in dieser Arbeit Beispiele die in letzter Zeit auf diversen Konferenzen für Digital Humanities oder Datenvisualisierung veröffentlicht wurden. Da das Fachgebiet Digital Humanities stetig an Popularität gewinnt werden die einzelnen Beispiele ihrer Geisteswisschaft zugeordnet und weiterhin in die benutzten Methoden unterteilt.

Contents

1. Introduction

Digital Humanities. Digital humanities is the intersection between traditional humanities and computational methods. Digital humanities developed with the use of computers in general but their application range has changed over time. While at first, it was used for digitizing archives and compiling databases, other computational techniques have been added: computer vision, data mining on large sets of data, computational text analysis, pattern recognition in images etc. Visualization is one of the computer-based methods that have been applied to data by humanities scholars. This work focuses on how networks are used in humanities as tools for data visualization.

Network Visualization. The topic of data visualization in humanities, was surveyed by Jaenicke et al. [Jan15b][Jä15] in 2013. While their work has its focus on the different visualization techniques that are used in the field of humanities and not only on data representing networks, this thesis intents to take a closer look on the topic of network visualization. For this, I did a research on various digital humanities and visualization conferences covering the past six years (ending in December 2016) to find examples for network visualization, namely in the *Digital Humanities* conference, the *Arts, Humanities, and Complex Networks* (AHCN) at NetSci, the *IEEEvis* conference and the *Eurovis* conference. I only searched six years into the past, because I felt that the results got much too scarce before that. Therefore, this survey is in no way exhaustive of the topic but shows more recent examples of network visualization applications in the humanities.
I found examples in the fields of history, archaeology, literature, art and musicology. Each of them uses different network visualization techniques and different types of networks. In the following, I present an overview of the types of networks I described in this survey.

Co-occurrence Networks. Co-occurrence networks are used to visualize potential connections between entities in a written source. Co-occurrence means that certain entities are mentioned in the same written article, book, chapter or word frame [Gra16b]. These entities are modeled as nodes and the edges indicate the existence of a co-occurence between the entities. Entities may be ficitonal characters of books or plays resulting in *character networks* but also people in historical documents like letters or research topics in scientific publications (e.g. name of genes [CHDS13]). In this thesis co-occurence networks are found in literary research and history.

Networks of Spatial Interaction. This network type uses a geographical layout and visualizes the interaction between the geographic locations. Thus, nodes represent locations on a map and edges, for example, represent how these locations interact if there is any flow of ideas or trading. These networks "are inherently spatial, with the flexibility to be both social and physical" [Bru10]. They can be found mostly in the field of archaeology.

Sentence-word Networks. Another type of graph visualises the successive affiliation of words and paths. Nodes here represent words. These words can belong to a set of words of a text or several texts. Edges in these networks connect words if they belong to the same subset of a set of words. This is used as a close reading tool in literature, for example to show the variants of text editions or rhyme schemes in a poem.

Networks of Attribute Similarity. This is found in art history, literature and musicology where data is stored in relational databases. Nodes are entities from a database, which could be artworks or musicians and the edges connect two entities if they share one or more descriptors or index terms or key words, attributes that are stored in the database. If the entities are persons, this can result in a *relationship graph* which is a social network and in a way similar to a co-occurrence network of people. Attributes could be origin, partnership, education, ect.

2. Preliminaries

2.1 Graphs and Graph Visualization

In this thesis we denote a *graph* G=(V,E) as a set of nodes V and a set of edges E. A graph can be viusualized in various ways but the most common type of visualization is to draw a graph as a *node-link diagram*. There are other types to visualize a graph, like the matrix method that is based on the adjaceny matrix.
The node-link diagram draws the nodes of a graph as a point on a plane and draws the edges as lines between the nodes that it connects.
The common goals of aesthetic visualization are: minimizing edge crossings and edge bends, and maximizing symmetry, edge orthogonality and node orthogonality [Pur02]. Another goal is that all edges have approximately the same length, and for directed graphs that the flow direction is consistent. This survey concentrates on these types of network visialization techniques.

Temporal Subnetworks. From any type of network, subnetworks can be derived. If the nodes in the subnetwork are attributes from a specific time frame it is called a temporal subnetwork.

In the following section, I will describe graph drawing algorithms that are used in the examples that are presented in this survey.

2.2 Algorithms

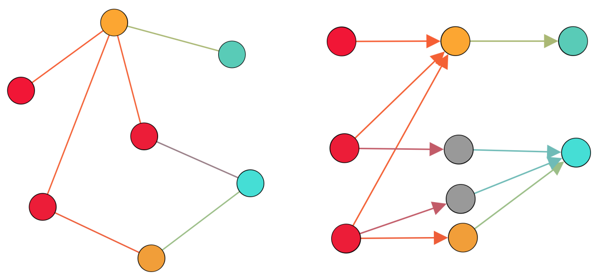
2.2.1 Force-directed Algorithms

With a *force-directed* layout, drawings of simple undirected graphs can be calculated. The basic idea is to consider a graph's nodes as physical objects and its edges as a set of physical springs and then let this system find a state of minimal energy [Kob13]. The force-directed algorithm of Eades [Ead84] formulated the spring force between two adjacent nodes as:

$$f_{spring} = c_{spring} \times log(\frac{d}{l})$$

, where c_{spring} is a constant. Non-adjacent nodes repel each other which is formulated as a repelling force:

$$f_{rep} = \frac{c_{rep}}{d^2}$$



(a) A force-directed drawing. (b) A layered graph example.

Figure 2.1: Two drawings of the same graph. Fig. 2.1a was created using the Fruchterman and Reingold algorithm; Fig. 2.1b is drawn as a layered graph (Red represents the first layer, orange the second layer, and blue the third layer. The gray nodes represent dummy nodes). Both graphs were created using the Gephi software.

where c_{rep} and l are constants. d is the distance between the nodes [Kob13]. The force-directed algorithm is then described as follows [Ead84]:

Algorithm 2.1: Algorithm of Eades

```
Data: Graph G
place nodes in random location;
repeat
    calculate the force of each node;
    move the node to c4∗ (force on node)
until M times;
draw graph on CRT or plotter.
```

The algorithm of Fruchterman and Reingold is an enhancement to the algorithm of Eades, where they added an even distribution of the nodes and added attractive and repulsive forces to nodes [FR91].

Graph Theoretic Distances. The method of Kamada and Kawai [KK89] tries to achive that the geometric distance between two nodes correspond to the graph theoretic distances between these nodes in the graph. The pairwise graph theoretic distances are computed by an *all-pairs shortest path* algorithm.

Large Graphs. All of the force-directed methods above only work well on small sets of nodes. The first algorithm for graphs of more than 1000 nodes came from Hadany and Harel [HH99], where the so called multiscale technique was used. For this technique, in the first step a fine-scale relocation is performed on the nodes. For this again, calculations from Fruchterman-Reingold [FR91] or Kamada-Kawai are used. In the second step a coarse-scale relocation is performed, again using calculations from Kamada-Kawai. In the third step a fine-scale relocation is performed.

2.2.2 Layered Drawings

Any directed graph (*digraph*) can be drawn in a layered fashion, which was first presented by Sugyama et al. in 1981 [Sug81]. With this drawing style, the vertices are placed on vertical or horizontal parrallel lines (layers) and edges are represented as monotone curves or polylines connecting their end nodes. The information on which node lies on which layer

is sometimes provided as an input and sometimes has to be the output.
If the graph is not acyclic, the cycles need to be removed first. For an acyclic digraph, three steps are needed to compute a layered drawing.

Layer Assignment. In the case that the information on which layers the nodes lie is not provided, computing this information becomes the goal of the first step of the algorithm. There are three possible optimization criteria for this step: First, the height and width should be small. This can be achieved by the *Coffman-Graham-Layering* [CG72]. To facilitate the next steps of the algorithm, the layering should be *proper* meaning that the span of an edge should not be more than one layer. This can be achieved by inserting dummy nodes - nodes that replace intersections of edges with layers. Since the larger amount of dummy nodes increases the time complexity of the following steps, the number of dummy nodes becomes another optimization goal during the layer assignment [GKNV93].

Crossing Reduction. The goal of the second step is to draw a layered digraph with a small number of crossings. The underlying problem is the *two-layer crossing problem* - the problem, given a bipartite set of nodes, to arrange nodes in such a way that the crossing between the two layers is minimized - which is NP-complete. There are heuristic methods that try to solve it [EK86]. Two methods running in linear time are the barycenter method and the median method. The barycenter method determines the x-coordinate of u in the second set $x_2(u)$ as:

$$avg(u) = \frac{1}{deg(u)} \sum_{v \in N_u} x_1(v),$$

where N_u is the set of all neighbours of node u, and $x_1(u)$ is the x-coordinate of u in the first set. The median method determines $x_2(u)$ as

$$med(u) = x_1(v_{\lfloor j/2 \rfloor})$$

The position of the node is calculated based on the position of its neighbours. Most of the time, a hybrid method is used where [dBT99]:

1. The median method is used for the initial order. If two nodes have an equal median value, the barycenter method is used for these.
2. The adjacent-exchange method is used to refine the output.

Horizontal Coordinate Assignment. In this step the exact x-coordinates are assigned to the nodes. A possible goal of this step is to reduce the angle at the bends caused by the introduction of dummy nodes, e.g. to draw the edges as straight as possible, without changing the order resulting from the previous steps [dBT99].

2.2.3 Drawings of Clustered Graphs

Clustering a graph means grouping the nodes of a graph G into - not necessarily - disjoint subsets of G that are called *clusters* [BC01][Fen97]. Visualizations that use some kind of clustering are often easier to read and provide additional insight on the graph structure. For example, if different nodes belong to different institutions, clustering can help not only to show the individual connections of each node but also how the institutions interact with each other, as shown in Fig.3.1b in chapter 3.
There are different methods to compute a clustering for a graph using different optimization criteria. On the one hand, the criteria can be to divide a graph evenly into distinct clusters. On the other hand, one might want to explore the structure of a graph like cliques or connectivity properties.
In the network examples presented in this thesis, clustering methods are used in almost all surveyed humanity fields.

2.2.4 Edge Bundling

The basic idea of *edge bundling* is to bundle compatible edges to reduce the total amount of edges and declutter the graph for a better visualization.
For edge bundling, a straight-line node-link diagram is used as an input. In hierarchical edge bundling, edges are bundled by bending each edge, modeled as a B-spline, toward the polyline path defined by the available hierarchy. A geometrical approach has been proposed by Cu et al. [CZQ$^+$09].
In a force-directed edge bundling approach, Holten and van Wijk subdivide the edges into segments and use a linear, attracting spring force between each pair of consecutive subdivision points. Between pairs of corresponding subdivision points in interacting edges, an attracting electrostatic force is used. A next step is to decide if edges are compatible to each other for bundling [HvW09]. Lambert et al. have described another routing algorithm which discretizes the plane into regions and routes edges around the regions's boundaries. [LBA10]. This is used for rerouting edges in [McC15].

2.2.5 Metro Map

The Metro-Line Crossing Minimization problem (*MLCM*) has been stated as [ABKS10]

> "[...] the problem of drawing a set of simple paths along the edges of an embedded underlying graph G = (V,E) so that the total number of crossings among pairs of paths is minimized."

This matches the drawing of a metro map where the stations are represented by a node and a metro line is represented by a path through all the stations it serves. The undlerying graph has to be drawn nicely and the line set has to be added to the underlying graph with as few crossings as possible to ensure a better readability. Algorithms for solving the the MLCM were presented by Noellenburg [Noe10] which was used for the visualization of a poem under 5.9 in chapter 5.

2.3 Graph-drawing software

Different graph drawing software programs were used in the examples of this survey, namely Gephi[1], Graphviz[2] - these two are open source - and OmniGraffle [3]. They provide different graph drawing algorithms one can select, and some other visualization tools. Gephi also uses plug-ins so clustering algorithm options are available too.

[1] https://gephi.org/
[2] http://www.graphviz.org/
[3] https://www.omnigroup.com/omnigraffle

3. History

In this chapter, I survey visualization approaches applied to networks in the field of history. An overview of data visualization through networks in the field of history was already written by Martin Grandjean who also published two research papers that I found for this thesis. In these, he states that networks should not only be used to visualize results but as an active tool for research [Gra15].
Two of the network examples here are networks showing how groups of people interact with each other, that can be classified as co-occurence network. Another network shows a geospatial network about how different historic places were connected through trading. This kind of network is also found in the field of archaeology.

3.1 Force-directed Approaches

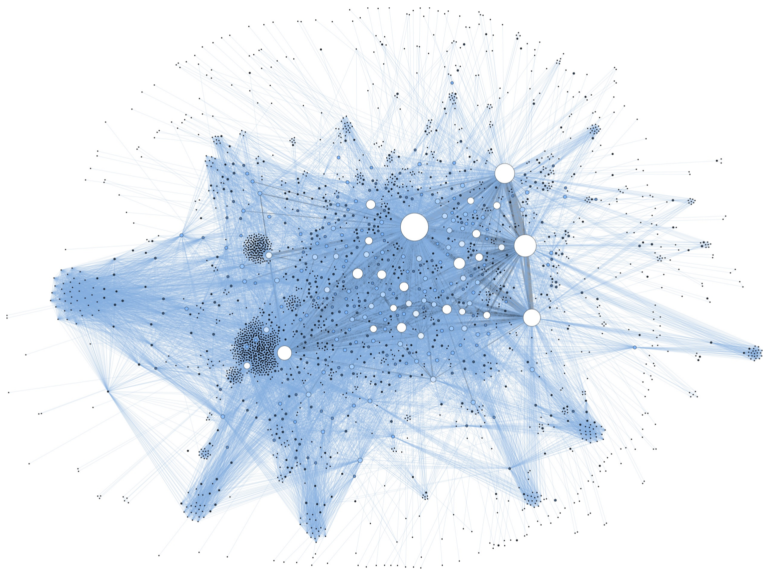
Archives Distant Reading: Mapping the Activity of the League of Nations' Intellectual Cooperation [Gra16a]

Data. This paper analyses the documents (kept at the United Nations Archive, Geneva) about the League of Nations, more precisely about the International Comittee of Intellectual Cooperation. They set up a relational database containing the documents and their metadata focusing on the people, referred to as *agents*, that are mentioned or adressed in the documents. 27,000 documents (university relations, bibliography coordination, educational matters and various enquiries) from the time between 1919 and 1927 were examined. The extracted database contains 3,200 different agents.

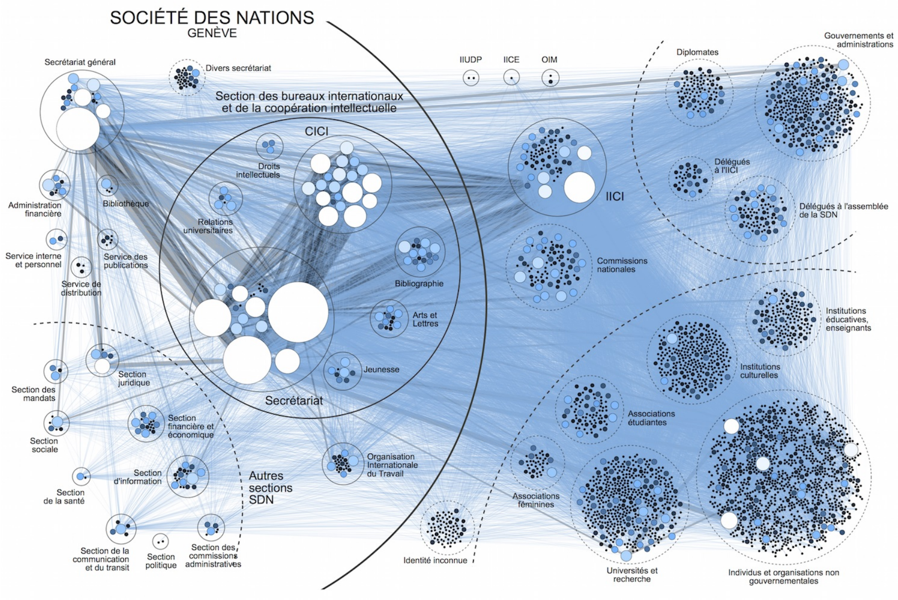
Graph. The graph can be categorized as a co-occurrence network where the nodes represent the agents and the edges represent a co-occurence between two agents in the same document. There are 3200 nodes in the graph and 38.600 edges are found between them.

Visualization. The two visualizations were created with the help of the Gephi software, presumably using a force-driven algorithm in Fig. 3.1a. In Fig. 3.1b a clustering was used in addition, probably created manually using the institutions.
In Fig. 3.1a, the size of a node is proportional to the number of appearances - in the data corpus - of the agent that it represents. The colour also indicates the number of appearances and varies between black if it appears under 10 times, blue if it appears between 10 and 50

times, and white if it appears more than 50 times [Gra14]. Each node was assigned the agent's name that it represents but that name it is not displayed in the examples.
The thickness of an edge is proportional to the number of co-occurrences between the connected actors. The colour also addresses co-occurrences and varies between blue if the count is less than 10 and black if it is more than 10 [Gra14].
Because of the large number of edges the nodes are grouped by the institutions to which the agents belong to in Fig. 3.1b, to show whether the institutions interact with each other. To improve the readability of interactions between the institutions a second layer of metadata was added to decribe the clusters. The clusters are emphasized by circles around them.



(a) Visualization using Gephi; a force-directed approach



(b) Network where all agents are grouped by their instiutions

Figure 3.1: The two different ways to visualize the network of agents in the League of Nations' Intellectual Cooperation; both figures are copied from [Gra15].

Protestant Letter Networks in the Reign of Mary I: A Quantitative Approach [Ahn13]

Data. The source of the information visualized here is literature about the protestant community in England during the reign of Mary I. The letters were taken from three books (John Foxes' *Acts and Monuments* and *Certain Most Godly, Fruitful, and Comfortable Letters of... True Saintes and Holy Martyrs of God*, and from Nicholas Ridley's *A pituous lamentation of the miserable estate of the churche of Christ in Englande*) and from a Victorian anthology. Only letters that had been send or recieved by residents of England, family correspondence or letters between Protestant fractions were accounted - amounting to a total of 289 letters.
The following information was extracted from the letters: the senders and receivers, their respective locations, the dates they were written and commendations or reported contacts.

Graph. A co-occurence network is constructed of the protestant members that are mentioned in the letters. There is a total of 377 persons, sharing 795 edges. The nodes in the graph represent the individuals extracted from the letters and the edges represent the links between them that were sorted in different categories: letter links, implied links, messenger links, spousal links and other family links.

Visualization. The graph was drawn using OmniGraffle, presumably using the algorithm of Kamada and Kawai.
Fig. 3.2 shows the main network of the global network of all 377 persons, which is shown in Fig. 3.3a.
An example for a comparison of temporal subnetworks here is the immediate environment of John Bradford - represented by the thick node in the center of the subnetwork. One subnetwork showing the network state right before his death and one for showing the state right after his death are compared in Fig. 3.3b. This shows how important certain individuals were for the coherence of the protestant community.

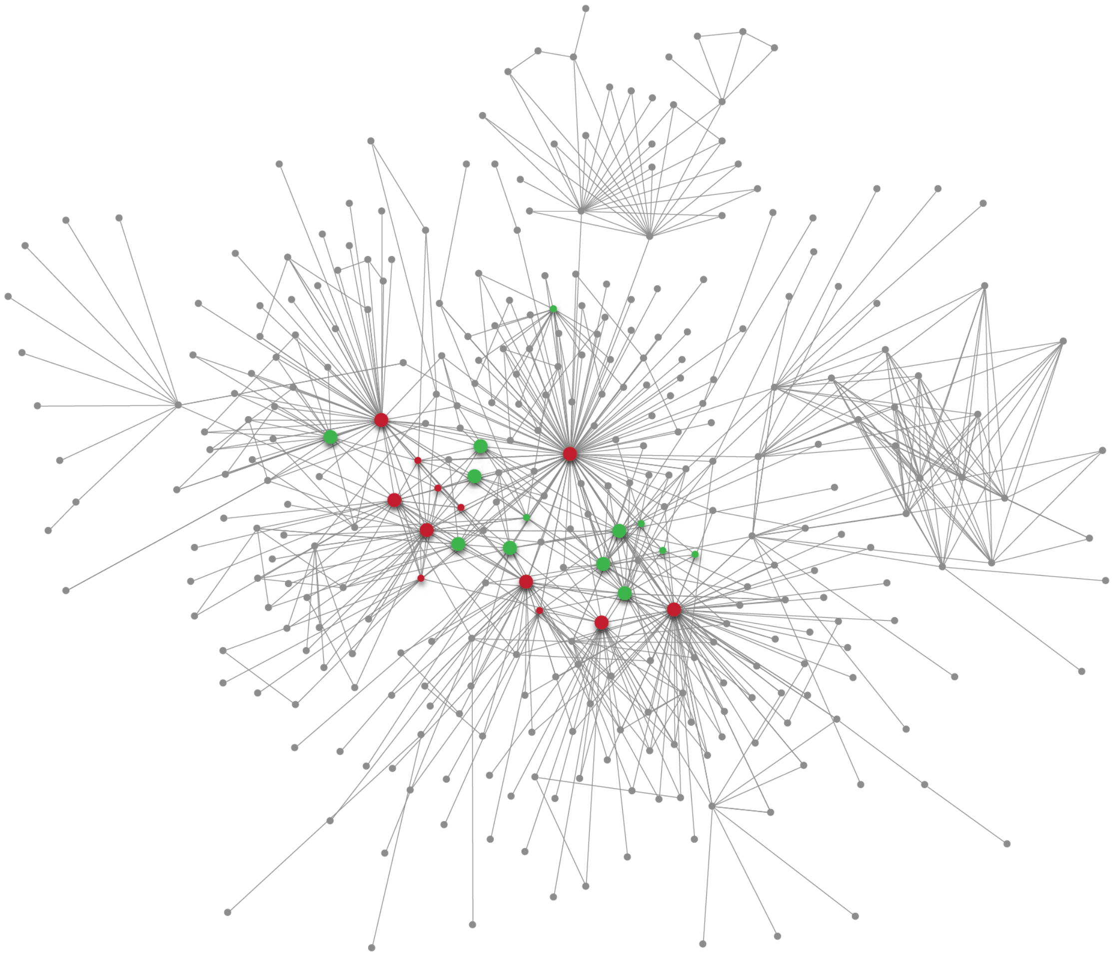
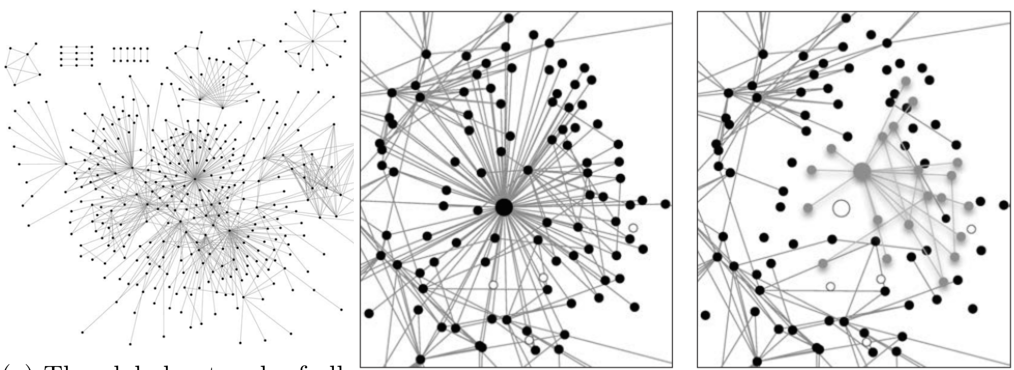


Figure 3.2: The main graph of the global network; copied from the book of abstracts of the AHCN of 2013 found on `http://artshumanities.netsci2013.net/`- the red nodes show prominent martyrs, the green nodes show prominent sustainers.



(a) The global network of all 377 participants and their social interactions.

(b) The subnetwork showing the immediate environment of John Bradford before and after his death.

Figure 3.3: The global network of all protestants and the subnetwork around John Bradford; copied from [Ahn13].

3.2 Geographical Approaches

A Network Analysis Approach of the Venetian Incanto System [Roc14]

Data. The Venetian Incanto system was an auction system to assign ship space on shipping lines, for the Venetian trading routes in the Mediterranian Sea. The data was extracted from official Venetian administrative documents.
After the documentes were digitized, transcribed and structured, 1480 lines of data were extracted. Each line of data resembles a shipping line and stores the line's name, the year, the number of ships, stopovers and duration of stay (optional). The period that was taken into account starts at the end of the 13th century and ends 1453.

Graph. The nodes represent ports existing during the 170 years in question, a directed edge from A to B represents maritime traffic from A to B through a galley line.

Visualization. There is one global network, shown in Fig. 3.4, that represents all the maritime traffic extracted from the data. In Fig. 3.5 temporal subnetworks were derived from the global network, each of them over the span of a year. For each node, the name of the represented port is shown next to it. The size of the node is proportional to the sum of its in- and out-degree measures. The edges have two attributes: the year and the number of ships owned by the line on this route.
The temporal subnetworks are important for the visualization because several temporal networks over a period of time visualize the development of the maritime traffic - the example in Fig. **??** shows how the Chioggia war, a war between Venice and Genoa from 1378 to 1381, affected the maritime traffic. The nodes are drawn in a spatial manner so that their positions and the distances between the nodes roughly resemble their positions on a map; it is possible that the drawing was created using some kind of constrained force-directed algorithm.

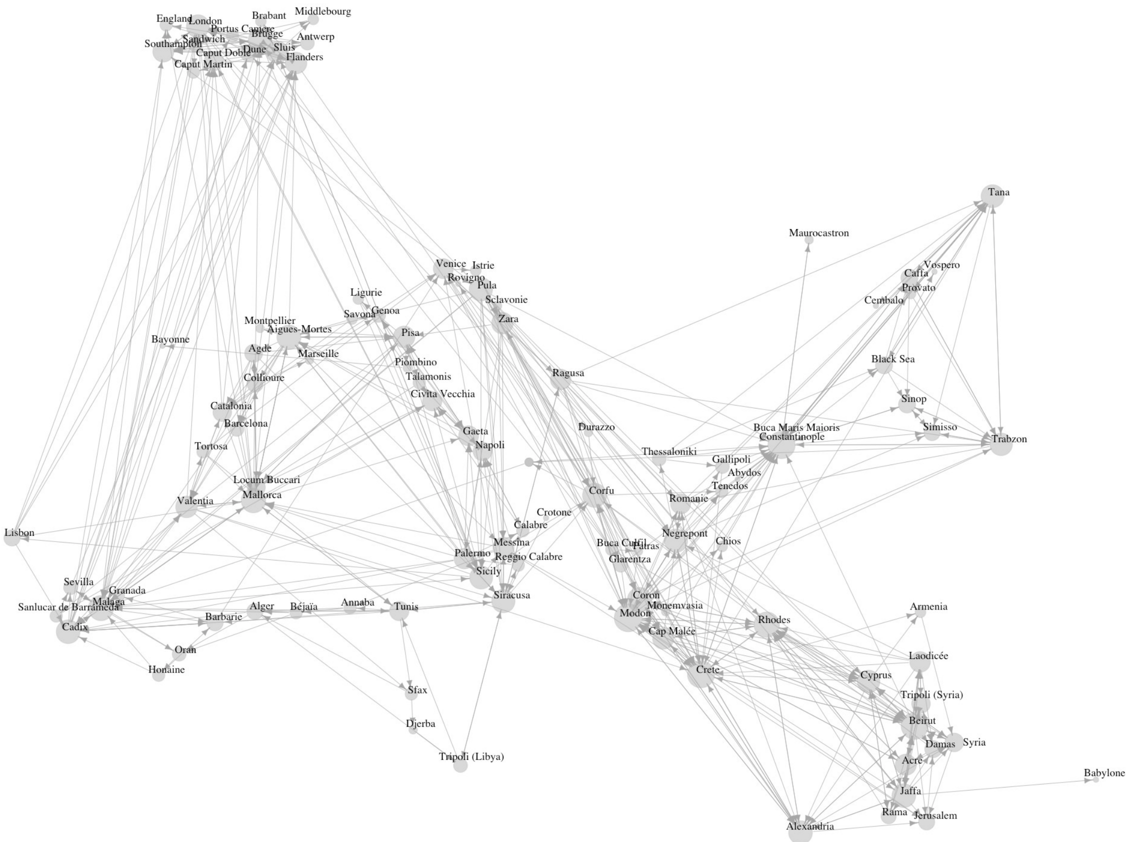


Figure 3.4: The global network presenting the Venetian Incanto routes over a span of 170 years; copied from [Roc14].

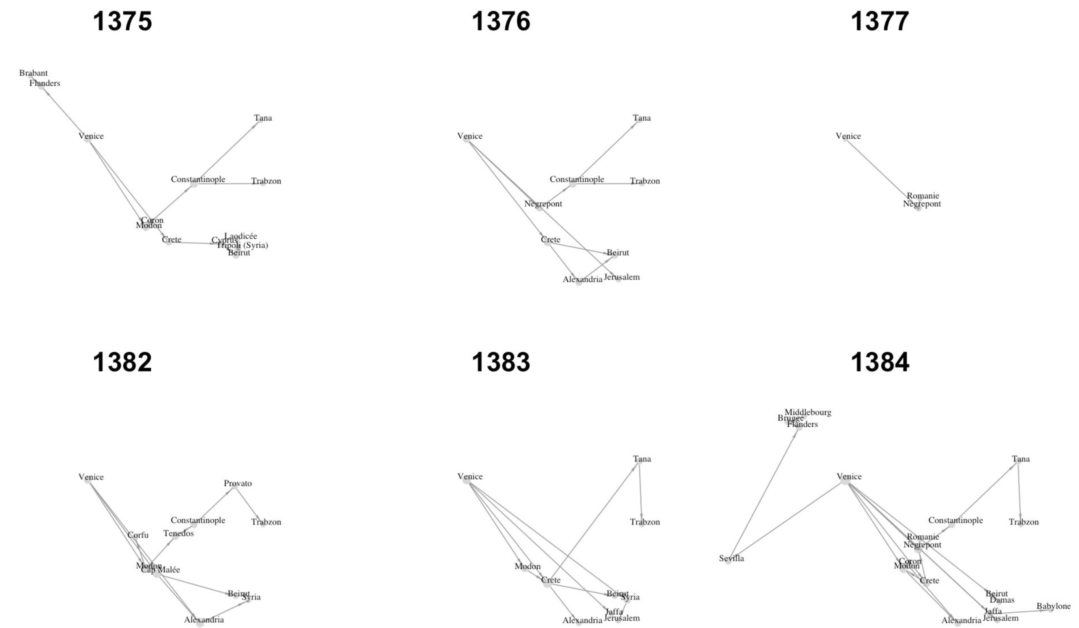


Figure 3.5: Temporal Subnetworks visualizing the traffic before and after the Chioggia war; copied from [Roc14].

4. Archaeology

In this chapter I survey visualization approaches applied to networks in the field of archaeology. Networks are a popular mean of data representation in archaeology.
Generally one could say that spatial interaction of settlements is more important in the field of archaeology than the interaction of specific persons. For that reason the networks focus on settlements or historic sites more than on individuals, and draw the maps in a geographic manner. The field of archaeology is the only field that I did not find any examples of force-directed layouts.
Application of social network analysis methods to these networks have been deeply studied by Tom Brughmans [Bru13][Bru10].

4.1 Geographical Approaches

Modelling Maritime Interaction in the Aegean Bronze Age [Kna08]

Data. The topic of this research is the interaction of 34 historic sites - e.g. cities or settlements - in the Aegean Middle Bronze Age. The sites were chosen because they were considered to be the most important settlements of the region at that time.
Each site is identified through the label i counting from 1 to 34. Each site is assigned its *local resource* value - local recources are tangible or intangible goods that a historic site possesses - and a variable that denotes its relative importance. A *physical separation* value and a relational *strength* were assigned to the links between sites. The relative importance and the relational strength are determined through optimizing a social cost-benefit potential function.

Graph. The graph consists of nodes that represent the 34 historic sites, and edges that represent the inter-site links - based on given data - between them.

Visualization. In Fig. 4.1, nodes are placed at their geographic location. Radiuses of nodes and thicknesses of edges are calculated as stated before. In Fig. 4.2, the same graph is visualized using the force-directed algorithm of Kamada and Kawai. In Fig. 4.1 each node is placed at the geographical location of the site that it represents. The size of the nodes is proportional to the size of that site. The thickness and darkness of an inter-site link is proportional to the strength of interaction between the sites. Nodes are labeled with the number of the site. In Fig. 4.2, the geographic affiliation of some nodes is marked as well (e.g. the Greek mainland or Crete) for a better comparison to the geographic model that was presented in Fig. 4.1.

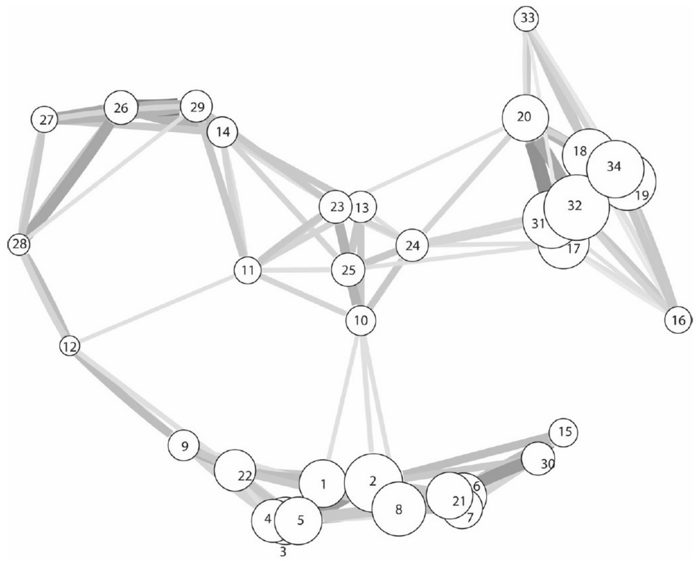


Figure 4.1: The proposed model of the Aegean Bronze Age; copied from [Kna08].

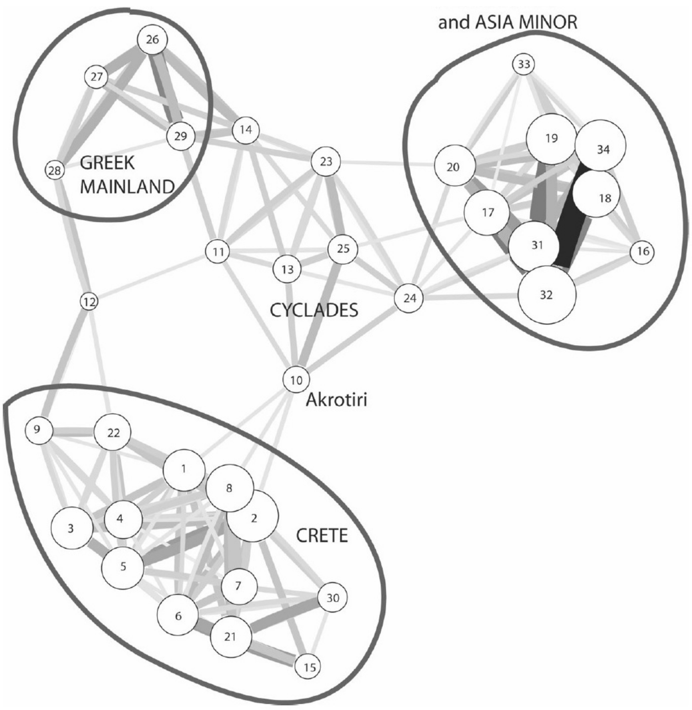


Figure 4.2: The same graph but the layout calculated with the algorithm of Kamada-Kawai; copied from [Kna08].

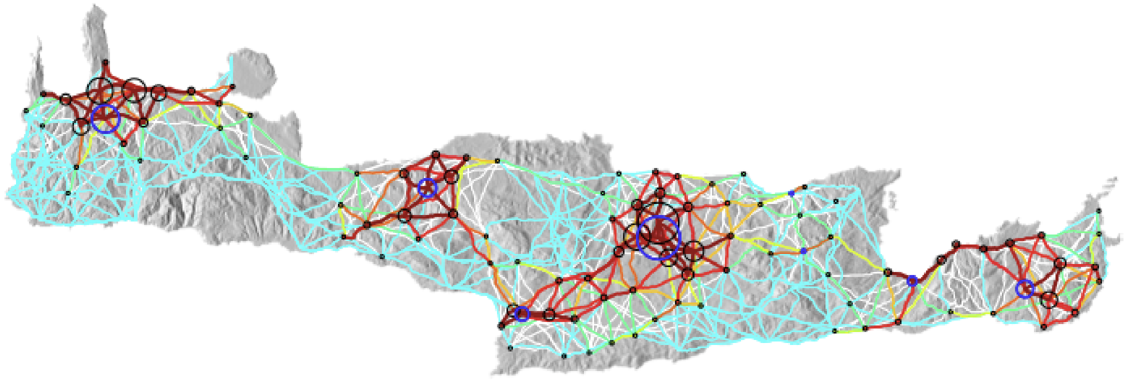


Figure 4.3: Example visualization of the settlement distribution of flow on the island of Crete; copied from [BW13].

Models of Settlement Hierarchy Based on Partial Evidence [BW13]

The topic of this research is Bronze Age settlement on the Greek island of Crete, including the political geography. Its special focus is on the interaction between the settlements, and also on the role of the road networks.

Data. Because most archaeological settlement distribution maps are incomplete, and therefore represent partial evidence only, in both importance of known settlements and entirely forgotten settlements, the existing data of the Crete settlements is extended by statistical methods, spatial analysis, and network techniques.

Graph. The nodes in the graph represent settlements, the edges represent realistic physical paths between them.

Visualization. Nodes are placed at their geographical position on a map. Edges are calculated through connecting each node to their nearest neighbours via shortest paths considering a digital elevation map of Crete.
To visualize a hierarchy via the settlements - to show the importance of the settlements - the settlements are marked by circles. The size of a circle is proportional to the relative importance of the settlement. Sites with an inflow total, that is larger than that of any site that it receives from are called terminal sites; their circles are shown in blue. Edges are coloured according to the amount of flow they hold. Blue, yellow and red represent low, medium, or high flow, respectively. White means no flow.

5. Literature

The most examples for networks in digital humanities, I found for the field of literature which is not surprising because in the field of literature the use of computational methods, and application of network analysis in particular, has been established by Franco Moretti [Mor05][Mor13] and are, even if the topic is still controversial, widely recognized in this field of study.
Many books have been analysed using social network analysis. Another option is the creation of character interaction networks - a more specialized co-occurrence network - can be created to visualize the social interactions in a book. These are mostly visualized using force-directed graph drawing algorithms, but there are also other examples. The interesting question here is how to define a interaction between two characters to link them in the network, based on the text.
I found two examples that used paths to visualize the membership of words to certain subsets of words from a text. These were realized as layered drawings.
The great benefit of literary works is that they can be processed using computational text processing. This means that, as opposed to the fields of musicology or art history, not only information about the works or their creators can be visualized, it is also possible to visualize the contents of the work itself in certain ways.

5.1 Force-directed Approaches

Character Network Analysis of Émile Zola's Les Rougon-Macquart [Roc15]

Data. The collection of books by Emile Zola, known as *Les Rougon-Macquart*, were published between 1871 and 1893. It consists of 20 novels.
An index for the whole series of books was used to create a table that contains the name of the characters, the names of the novels, and the pages in which they appear. From this table, they derive the co-occurrence networks after splitting the big table into smaller ones; one for each novel.

Graph. The resulting graph is another example for a co-occurrrence network. More precicely it is a collection of co-occurrence networks, character networks. One character network was created for each novel. First, bipartite networks were built from the extracted tables that linked a set of characters with a set of pages. From these two-mode networks, corresponding one-mode networks were created. Therefore an edge in the graph represents a co-occurrence on the same page of the novel. The number of nodes varies between 16 and 88 and the number of edges varies between 68 and 1181.

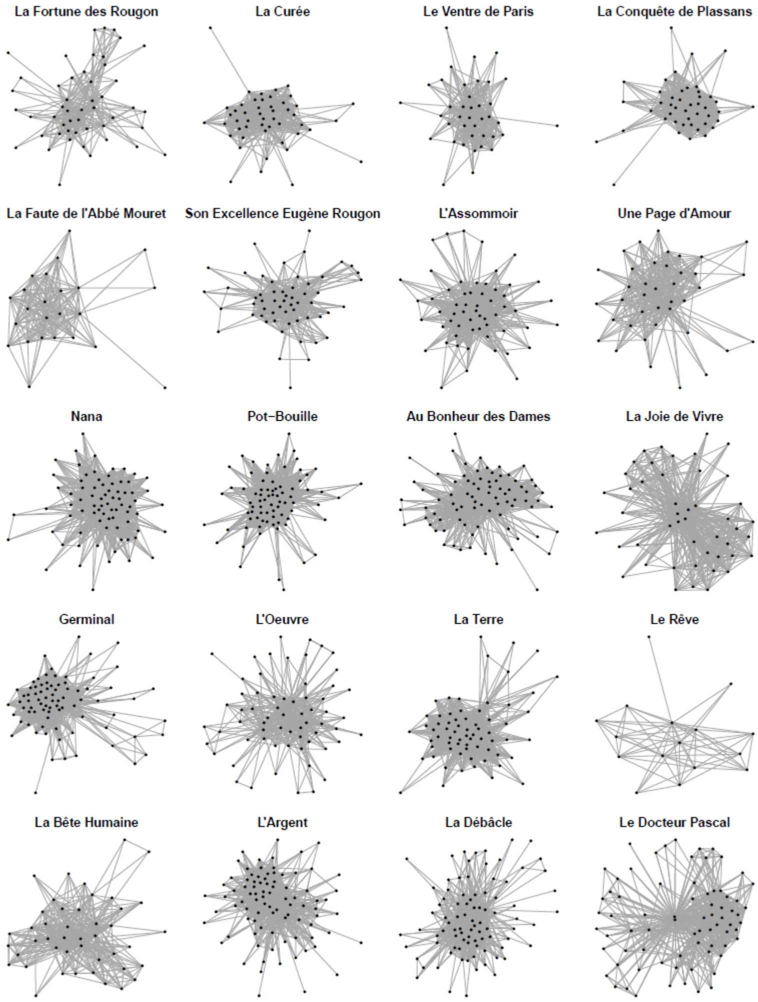


Figure 5.1: All networks of the twenty novels of the series *Les Rougon-Macquart*; copied from [Roc15].

Visualization. The force-directed algorithm of Fruchterman and Reingold was then used to calculate the visualization of the character networks which are shown in Fig. 5.1.

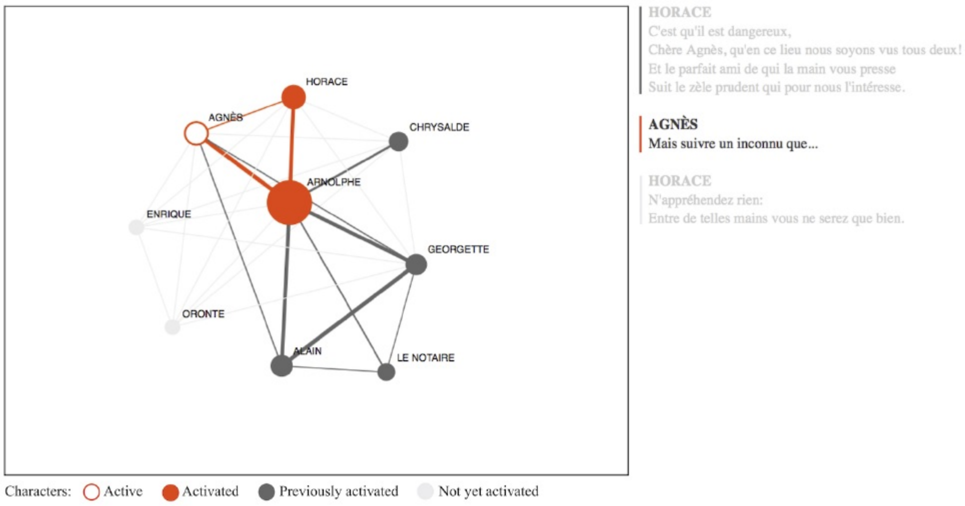


Figure 5.2: *L'ecole des femmes, act 5, scene 3; copied from [GRPX16].*

Visualizing the Dynamics of Character Networks [GRPX16]

This paper analyzes the character interaction in Moliere's play 'Ecole des femmes' from 1663.

Data. The text was provided as raw text from the Gutenberg library. To extract the structure of the play, Orange Textable (a python package to extract tables from raw text) was used. The important information was which characters were present in each scene and their respective lines.
The information was collected in two tables. They both have the same number of entries which corresponds to the number of lines in the play. The first table contains the speaker and the content of the line, the second table contains all the characters that were present in the respective scene.

Graph. From the tables, a co-occurrence network was constructed for each line. There are 9 nodes in total, one for each character.

Visualization. A force-based layout is used for visualization. The network state is visualized for each line. The nodes have different colours representing the statuses of the characters in the respective lines. The colour legend is blank for active (speaker), orange for activated (presence in the scene), dark grey is previously present and light grey is not yet activated.
The node's radius is proportional to the number of lines that were spoken by the character until then. It accumulates throughout the visualization and thus increases over time. The thickness of an edge is proportional to the number of co-occurences of two characters. It also increases over time. This is shown in Fig. 5.3.
For information about the actual line content, the line is displayed next to the graph, as shown in Fig. 5.2.

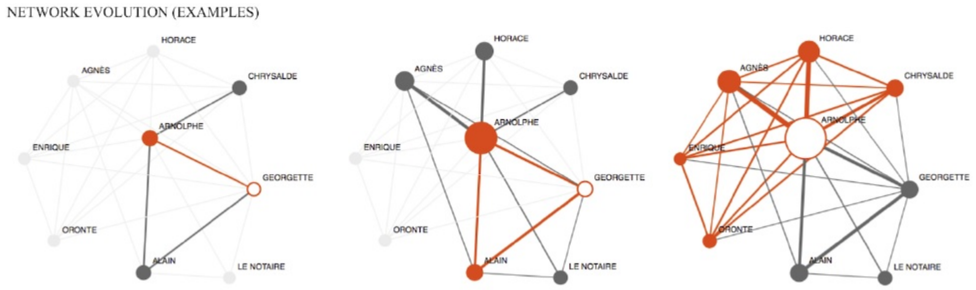


Figure 5.3: Three character networks from different stages of the play; copied from [GRPX16].

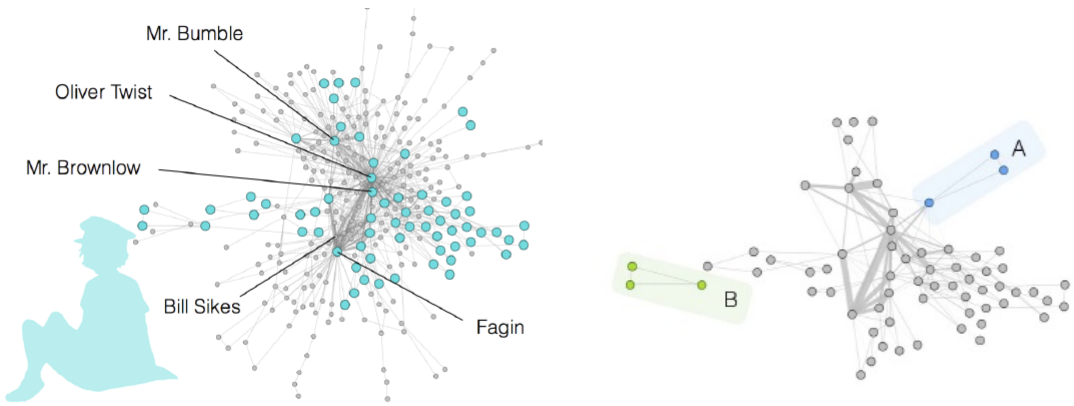
Discovering Structure in Social Networks of 19th Century Fiction [Gra16b]

This paper explores the structures of novels by Jane Austen and Charles Dickens with a focus on character interaction.

Data. Six novels by Jane Austen and three novels by Charles Dickens were examined in this work. The raw text version of the novels was obtained from the Project Gutenberg library.
For each node, a table of characters was extracted with one entry per character stating all their respective aliases. Then all aliases in the novel were replaced by the actual character's name. Thus, the co-occurrences could be extracted. A co-occurrence is defined between two characters in this novel if their names appear in a certain sliding window within a predefined number of words. Afterwards the co-occurrences for each pair where counted in each chapter.

Graph. A co-occurrence network is created for each chapter at first. Then, the graphs are aggregated to form a graph that represents the whole novel.
A node is created for each character in the character table.
An edge between two nodes means that there is at least one co-occurrence between the two characters in the chapter.

Visualization. It resembles a force-directed layout. The thickness of an edge is proportional to its weight which is proportional to the number of co-occurrences between the end nodes.
Communities in the graphs were detected by using the OSLOM algorithm (Order Statistics Local Optimization Method, an open source clustering algorithm for networks). The communities are then highlighted in a colour (in Fig. 5.4a, they are light blue). The nodes that don't belong to any of them are shown in grey.
Fig. 5.4b shows a subnetwork of the *Oliver Twist* global network, that contains all the nodes in communities. Two subnetworks are highlighted that are linked to *micro-narratives* - sub-narratives that only involve a small number of characters.



(a) The global network of *Oliver Twist*, communities are highlighted.

(b) Subnetwork of *Oliver Twist*, showing only the nodes that are included in a community.

Figure 5.4: Two graphs that visualize the communities in the novel *Oliver Twist*; copied from the poster presented at the websci 2016.

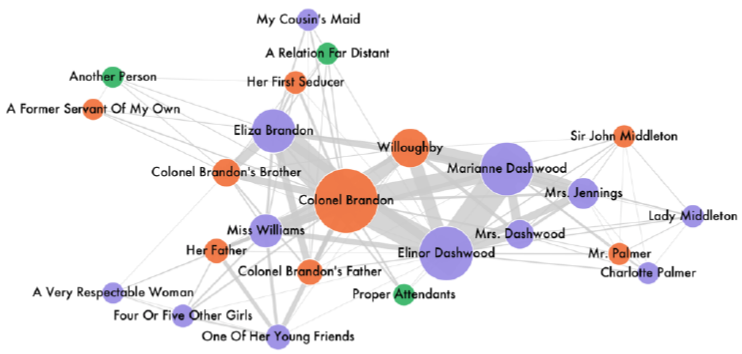


Figure 5.5: Network of chapter 31 from *Sense and Sensibility* using a combined strategy; copied from [Gra16c].

The Sense and Sensibility of Different Sliding Windows in Constructing Co-occurrence Networks from Literature [Gra16c]

The focus of this work is on the different definitions of co-occurrence in literature.

Data and Graph. The data and how the networks are constructed are the same as in the previous section, as in the co-occurrence networks for novels by Jane Austen. This work focuses on how the co-occurrences are defined in more detail.

Visualization. The graph resembles a force-directed layout.The size of a node is proportional to the node's weighted degree, the thickness of an edge is proportional to the number of co-occurences between the nodes that it connects.
Additionally, the character's gender is reflected through different colours: orange represents male charactes, purple female, and green represents a collective of people or a person of unknown gender.
Fig. 5.5 and Fig. 5.6 show the same chapter of Jane Austen's *Sense and Sensibility* with the same window size - just using different types of co-occurrences.

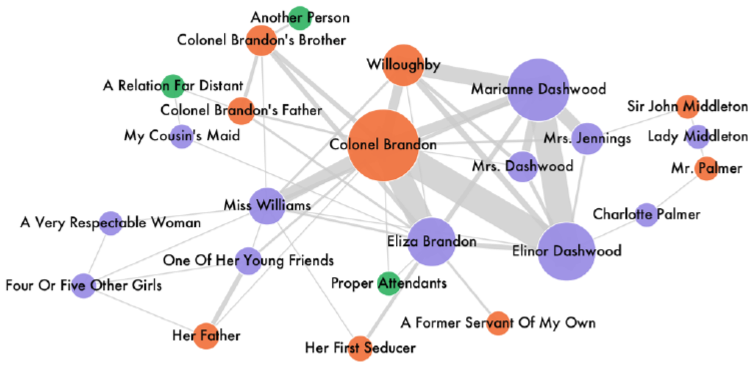


Figure 5.6: Network of chapter 31 from *Sense and Sensibility* using the collinear strategy; copied from [Gra16c].

5.2 Layered approaches

TRAViz: A Visualization for Variant Graphs [JGS16]

TRAViz is short for *Text Re-Use Alignment Visualization*. It is used to visualize how different versions or translations of a text vary.

Data. Books that existed in different versions in raw text format were used here, in this example several English bible editions and German translations of Shakespeare.

Graph. The graph is a directed acyclic graph. It is read from left to right in the same manner as we would read the text. The nodes represent words. Additionally there exists a 'super source' node connected by an outgoing edge to the first word of each edition and a 'super sink' node connected by an incoming edge to the last word of each edition. This graph is called a Variant Graph. An edge connects two consecutive words of the same edition. Thus a single edition forms a path.

Visualization. The algorithm that is used here is a variant of the Sugyama framework. Additionally, the edges of the layered graph are bundled. The design rules for visualization are the following:

- Words that are versions of each other in different editions of the text are shown on the same vertical line, i.e. on the same layer.
- A distinct colour is assigned to each edition's path.
- The label (word) size is proportional to the occurrence of a word across all the editions.
- The are no arrows on the edges because it is a layered drawing and the reading direction is set.
- Edges are not labeled as each edge in each edition is already drawn in a different colour - if there are more than 10 editions different colour hues are used.
- Major edges are bundled in thick gray strokes.
- Line breaks are inserted as necessary (depending on the width of the browser window) to improve reading, as shown in Fig. 5.7.

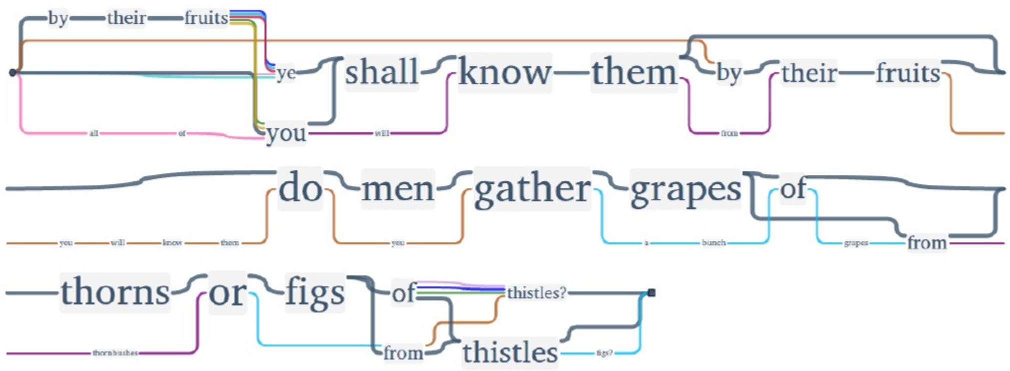


Figure 5.7: Different English editions of the bible with line breaks; copied from [JGS16]

Poemage: Visualizing the Sonic Topology of a Poem [McC15]

Data. The poems are provided as raw text. Fourteen poems are already included by default to the software but custom poems can just be added in raw text. The poemspace (the 2D representation of a poem regarding whitespaces etc), rhyme sets (sets of words linked by a rhyming scheme) and the sonic topology (the distribution of rhyme sets across a poem, respectively their interaction) are analyzed.

Graph. A node represents a word in the poem at its corresponding location in the poemspace. Edges connect two consecutive words in a rhyme set, thus a rhyme set is represented by a path similar to the TRAViz visualizaton.

Visualization. There are three views (see Fig. 5.9):

- The *set view* lists all rhyme sets that are found in the poem. They are shown as circles. If you mouse over the circles the letters or syllables they represent are displayed. The size of the circles is proportional to the relative number of words that the rhyme set contains. If a rhyme set is selected, a colour is assigned to it.
- The *poem view* shows the poem in its text form. If rhyme sets are selected in the set view, the participating words are circled in the assigned colour.
- The *path view* visualizes the paths of the selected rhyme sets. Each path

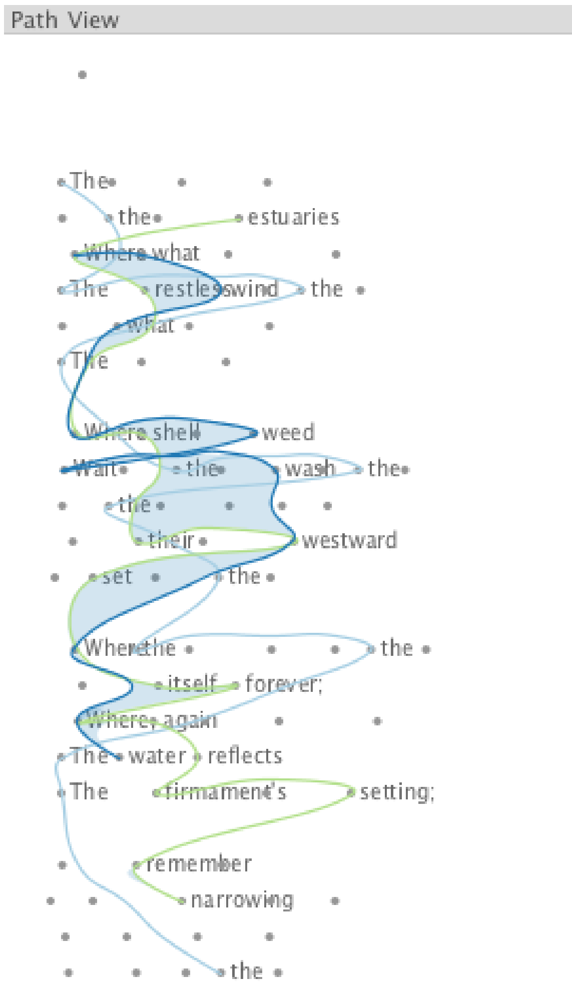


Figure 5.8: The path view of Louise Bogan's *The Night*. Three rhyme sets are selected for the visualization.

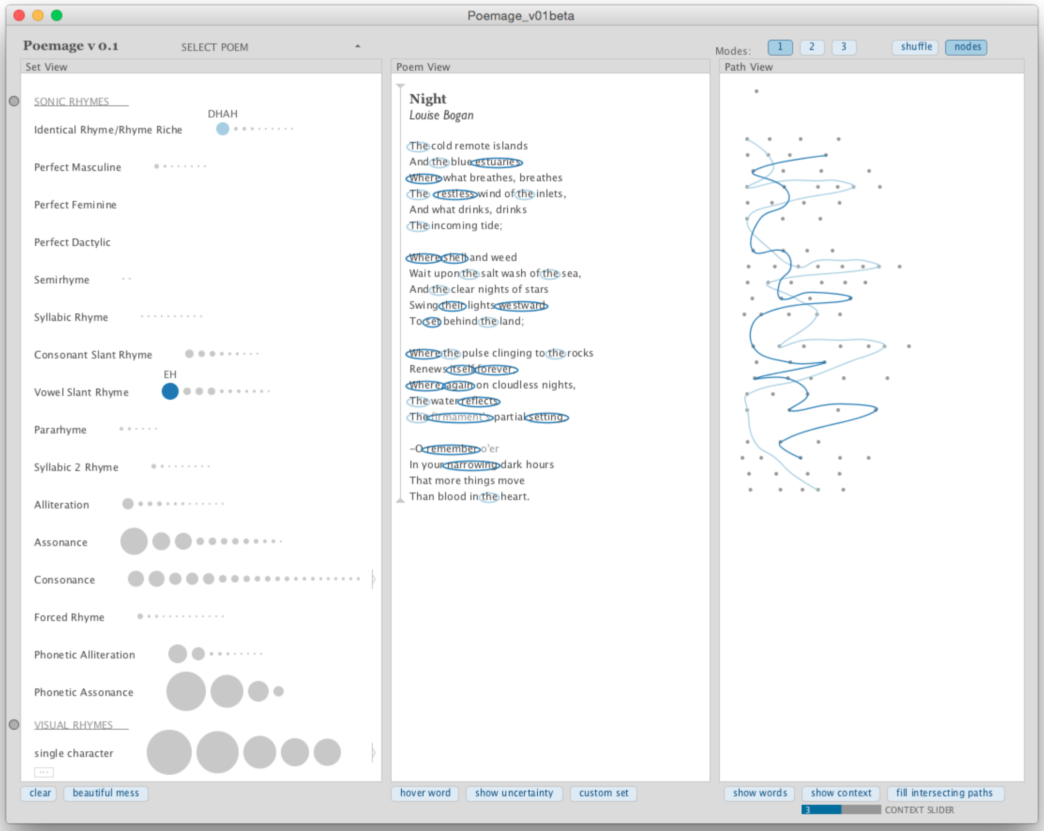


Figure 5.9: The three views of the visualization, left: set view, middle: poem view, right: path view

has the assigned colour of the visualized rhyme set. Edges are rerouted so they exclusively contain the nodes of the rhyme set.

The paths are rerouted to avoid ambiguous memberships of words. To draw multiple paths - which resembles the MLCM - they adapt an approach form Noellenburg [KK89]. Edges that connect words that are separated by more than one layer in the poem are rerouted in a way that the pairwise orders are maintained by pairs of paths sharing common subpaths.

5.3 Manual Drawings

Extracting Social Networks from Literary Fiction [Els10]

Data. 60 nineteenth-century novels from 31 authors were analyzed for this. The used books were provided as raw text.
First, all characters were extracted from the books including their aliases, or the way they could be adressed. All instances of quoted speech were marked as well. Then, a speaker (from the characters) was assigned to each instance of quoted speech.

Graph. One graph was extracted for each of the 60 novels. The graph shows a specialized co-occurrence network, described as a *conversational network*. Therefore the nodes represent

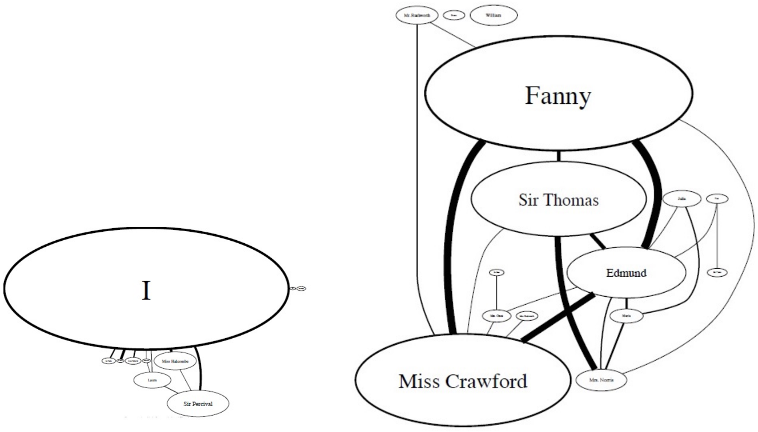


Figure 5.10: Collin's *The Woman in White* and Jane Austens's *Mansfield Park*; copied from [Els10]

the characters of the novel and the edges represent co-occurrences defined as dialogue interactions between two characters. In a node, all the aliases or mentionings of a character are collected.

Visualization. The networks were built manually. First, a node was assigned to each character (only characters that were mentioned more than three times in the text), and their aliases were grouped in the same node. Undirected edges were drawn between nodes that were connected through quoted speech. This connection means that quoted speech from each character falls within 300 words of one another with no quotes from another character inbetween.
The weight of an edge is proporional to the total length of words, from all the quotes that either one of the characters speaks in the quoted speech that connects them. All weighted edges are normalized by the length of the novel. The weight of an edge is proportional to the number of interactions between the two characters by which they are connected. The weight of the nodes is not mentioned, but it seems to be proportional to the frequency that the character is present in a dialogue.

5.4 Bimodal networks

„im Zentrum eines Netzes [...] geistiger Fäden“. Erschließung und Erforschung thematischer Zusammenhänge in heterogenen Briefkorpora [Hil15]

Data. The letters that were analyzed in this work were written in the context of the formation of a newspaper, *Die Sammlung* (the collection) by Klaus Mann, that first came out in 1933.
This research project focuses on the indexing of the texts and the extraction of index terms and keywords. The extraction of themes was accomplished by first constructing a thesaurus for index words, then analyzing the preprocessed texts for keywords using *Part-of-Speech Tagger* - a tool used to assign the words and punctuation to the part of speech e.g. identify the nouns, verbs, etc. Finally the keywords of a letter are assigned to the index words of the thesaurus and the index words to each other.

Graph. The network in Fig. 5.11 displays the mapping of index terms, that is how different letters share index terms, referred to as a *thematic network*.
The graph is realized as a bimodal network and can be classified as a network of attribute similarity. It is a bipartite graph with one part of the nodes representing letters - which are endcoded *b-00XX* - and the other one representing index terms. Edges between index terms and letters are drawn if a term appears in at least one other letter.

Visualization. The font size of an index term is proportional to the number of letters it appears in. I assume that the same applies to the font size of the letters.
The nodes that represent letters are coloured blue; the nodes that represent index terms are coloured purple.

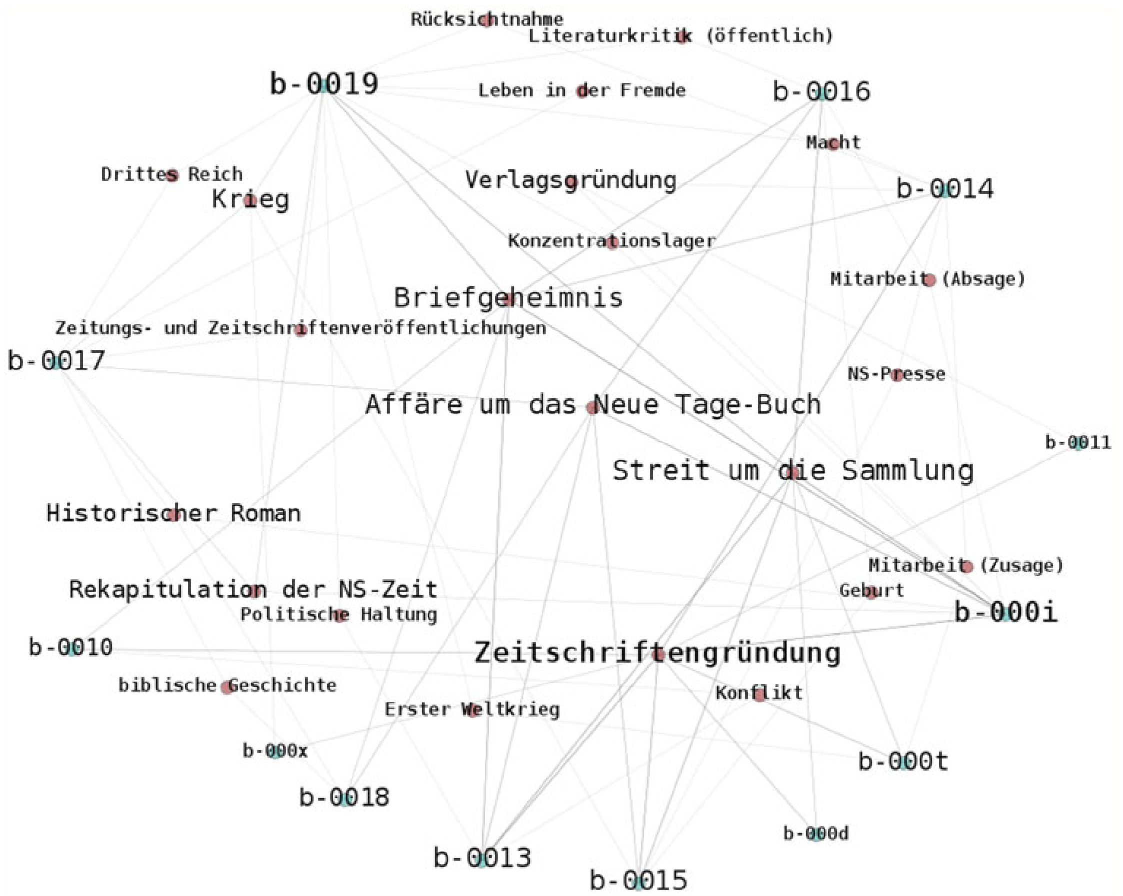


Figure 5.11: The *thematic network* extracted from letters around *Die Sammlung*; copied from [Hil15]

6. Art History

The next chapter is about the application of network visualization techniques in the field of art or rather art history. This chapter is rather small as it was difficult to find many examples for this field of digital humanities. Therefore it is hard to make general assumptions about the network visualization techniques used here.
The examples that I found are based on relational databases that contained information about artworks and artists. Here again, social network analysis is used to analyze the interactions between artists and how they influenced one another.
The networks that are constructed from the databases are visualized using force-directed graph drawing algorithms and, in one example, additional clustering techniques, and can be classified as networks of attribute similarity. The use of these networks is to see how characteristics of artworks are distributed and how ideas may flow through the community of artists.

6.1 Force-directed approaches

Towards a Digital Geography of Hispanic Baroque Art [Sua12]

Data. The basis of this work is the Baroque Art Database[1]. It contains more than 12,000 paintings by more than 1,500 creators, originating from the era of the Hispanic Monarchy in the period between the 16th and 19th century. The items in the database are characterized by around 200 *descpriptors* - these are tags that describe the artwork or the topics of it, e.g 'saint', 'virgin', 'christ', etc.
This piece of work contains *manual semantic annotations* of the artworks together with the descriptors, with an average number of 6 descriptors per artwork. For the network, only artworks with six or more descriptors were used.

Graph. There are two types of graphs in this work: one can be described as an *artwork graph* and the other as an *artist graph.* In the artwork graph the nodes represent artworks and an edge is drawn between two nodes if they share at least one descriptor. The weights of the edges are derived from the number of shared descriptors. In the artist graph the nodes represent artists. Therefore, the edges represent some sort of relation between the artists. I assume that they are also extracted from the descriptors of their artworks - to see how they influence each other's work. Both networks can be classified as networks of attribute similarity, and the artist network more specifically resembles a relationship graph. In Fig. 6.2, the nodes represent artists.

[1] `http://baroqueart.cultureplex.ca/`

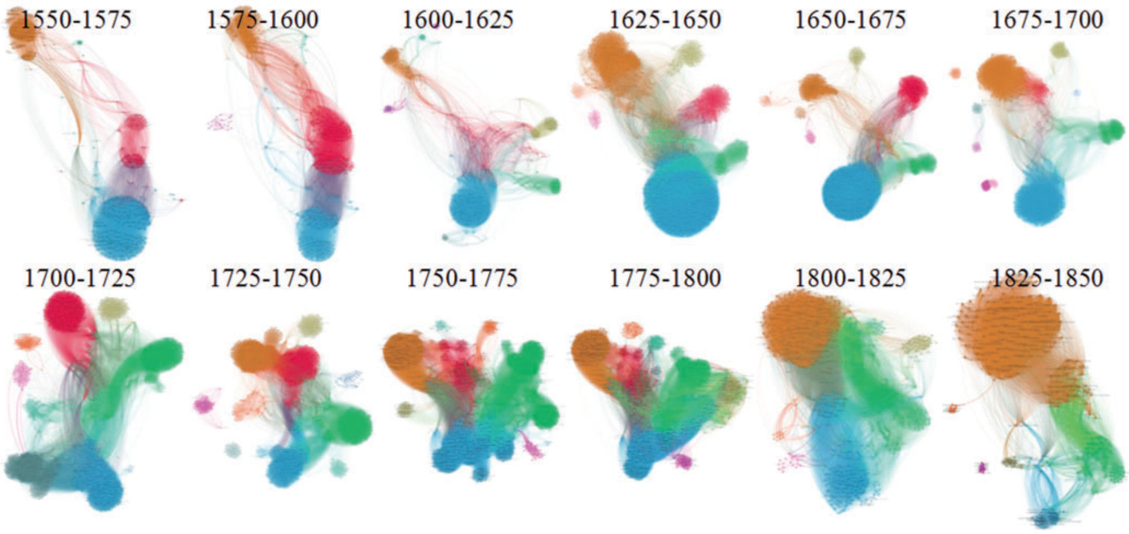


Figure 6.1: Clusters of artworks from 1650 to 1850; copied from [Sua12]

Visualization. In Fig. 6.1, twelve temporal subnetworks of the artwork graph are visualized, each covering a period of 25 years. For each period, the clustering classes are determined, and *distances* between them - measured by the frequency of descriptor usage - are calculated. These classes are distributed in a two-dimensional space so that their relative positions represent the relative distances among them. Then, a potential field for each descriptor was generated and the distances between similarity classes in different periods were calculated. This is to show the evolution of similarity clusters. The size of a cluster is proportional to the number of artworks.
Fig. 6.2 visualizes the artist graph. The higher the percentage of artworks produced by an artist, the larger is the font size of their name.
In both visualizations, circular arcs are used for edges.

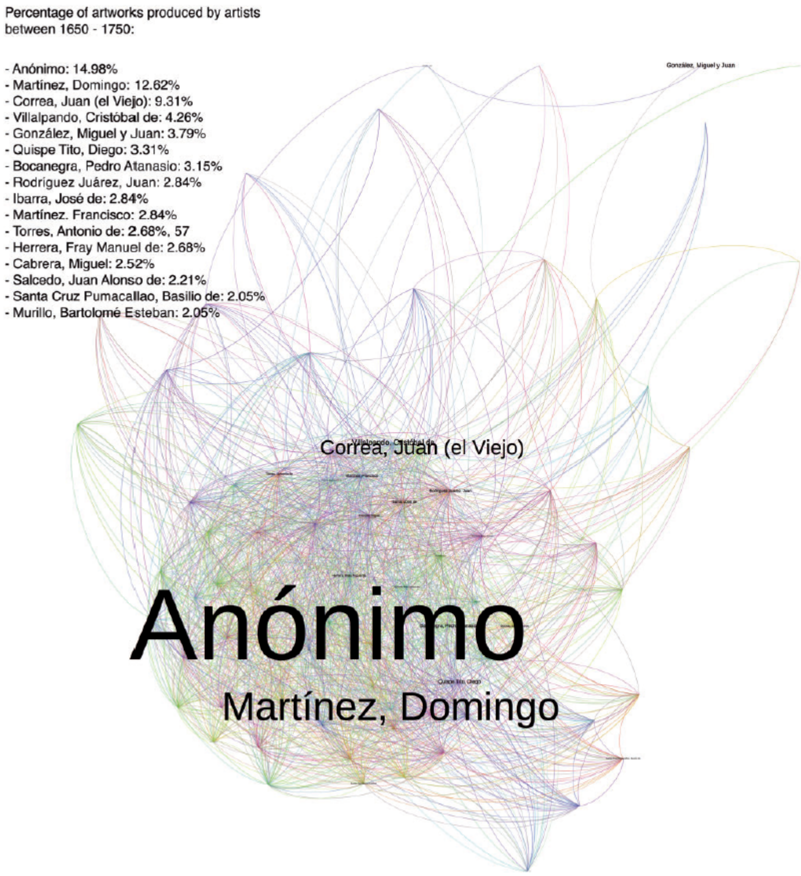


Figure 6.2: Network of artists; copied from [Sua12]

MapTap and Cornelia: Slow Digital Art History and Formal Art Historical Social Network Research [Bro16]

This work is about the tapestry history and especially the role of women in the tapestry universe.

Data. The base of this work is the art database *Cornelia*. The data stored in Cornelia consists of parish records, registers of guilds and corporations, district books and notarial deeds, such as business contracts, marriage agreements and credit transactions. For every source, Cornelia includes all attribution and relational data that connects actors - people acting in the tapestry history as artists, retailers, etc. - with other actors, places or works of art. By February 2016, it contained over 4300 actors.

Graph. The graph that was created from the data is a global network of all actors. The nodes therefore represent the actors and the edges between them represent connections, either family connections or business connections.

Visualization. The visualization resembles a force-directed approach. Fig. 6.3 and Fig. 6.4 are subnetworks of the global network, each focusing on the social environment of an actor. It is not expressed in the text, but I assume that the size of a node is proportional to the node's degree.
Fig. 6.3 shows the environment of Albert Auwercx who was a *tapissier* - an artist of tapestry. Nodes that represent women are purple coloured and nodes that represent men are lime coloured.
Fig. 6.4 shows the direct environment of Maria de Smit and Sara de Smit who were retailers of tapestry in Antwerpe. Nodes representing actors from Antwerpe are orange,

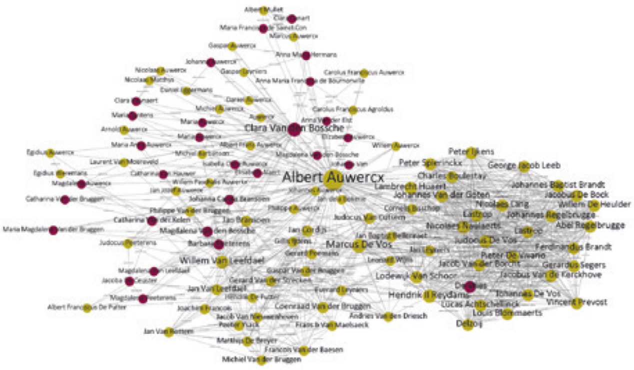


Figure 6.3: Subnetwork of Albert Auwercx, Clara van den Bossche (1650s - 1660s); copied from [Bro16]

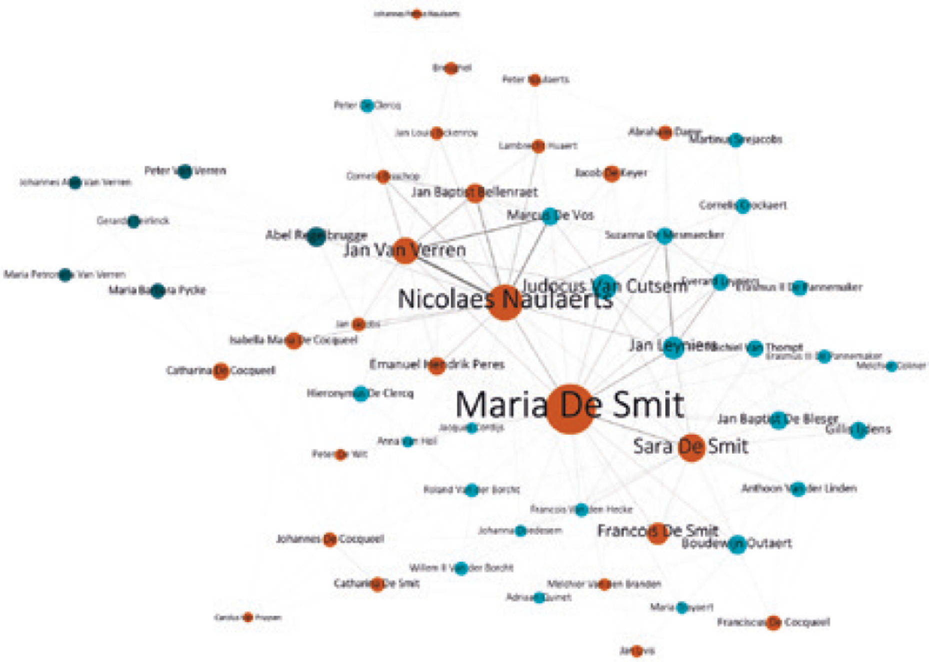


Figure 6.4: Subnetwork of Maria and Sara de Smit (1660s - 1670s); copied from [Bro16]

nodes representing actors from Brussels are light blue, and nodes representing actors from Oudenarde are dark blue.

7. Musicology

The results of network visualization techniques in the field of musicology are similarly scarce as in the field of art history, and the and the data is also accessed in similar ways using databases and comparing attributes - thus, the examples here are networks of attribute similarity.

7.1 Layered Approaches

Interactive Visual Profiling of Musicians (Part A) [Jan15a]

Data. This work is based on the *Bavarian Musicians Encyclopedia Online*[1] (German: "Bayerisches Musiker-Lexikon Online [BMLO]") . This encyclopedia contains the biographical information of 30,000 musicians.
For this work, the similarities between musicians are investigated, focusing on gender, time of activity, activity region, musical profession, further profession, relationships, division and denomination.

Visualization. The *Column Explorer* (Fig. 7.1) is a column-based representation based on list view in the Jigsaw framework and Parallel Tag Clouds. It visualizes the correlations among attributes. It can be viewed as a layered graph layout where a node represents an attribute in a column and an edge connects the coherent attributes that belong to each musician in each column. The columns are the levels of the graph. The x-coordinates are sorted according to the similarities from the selected musician m to musicians $s_1, ..., s_n$. The second column shows the encoded lifetime data. The next columns list the muscial professions, further professions, divisions and denominations of the listed musicians.

Each musician's name is highlighted with a specific colour - a mouse-over shows additional attributes. The font size of an attribute encodes the number of occurrences.
There is a horizontal stream for each musician that marks all the attributes that relate to the musician. It is drawn in the musician's assigned colour. In the second column the stream broadens proportionally to the span of the timeline's active time (first mentioning to death). In the next columns, the stream splits up if there is more than one attribute in a column.
The lifetime data is visualized as a vertical timeline, shown by the musician's colour.

[1] `http://www.bmlo.lmu.de/`

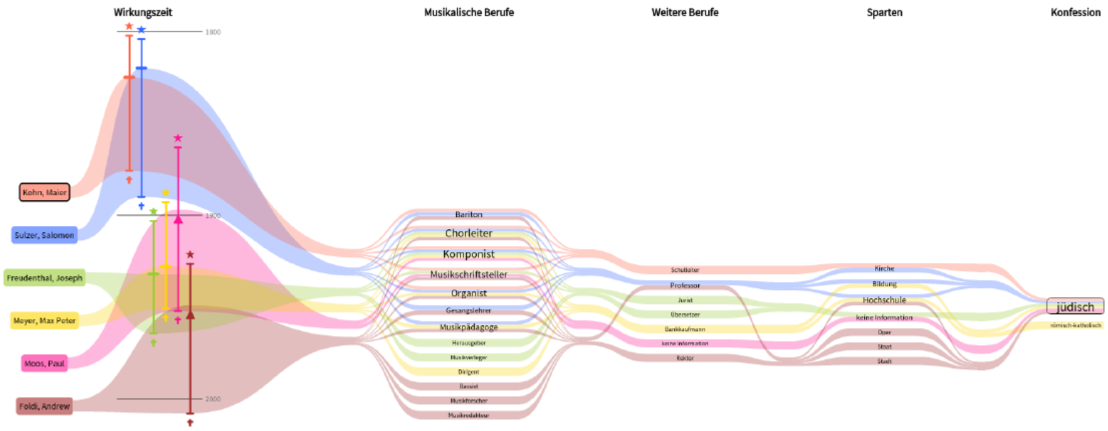


Figure 7.1: The visualization of the Column Explorer; copied from [Jan15a]

Different shapes on the line indicate different events, such as birth (★), death (a cross) and different signs (▲, ▼, •, ...) for the exactness of the dates.
The algorithm that is used for this visualization is not described in the publication.

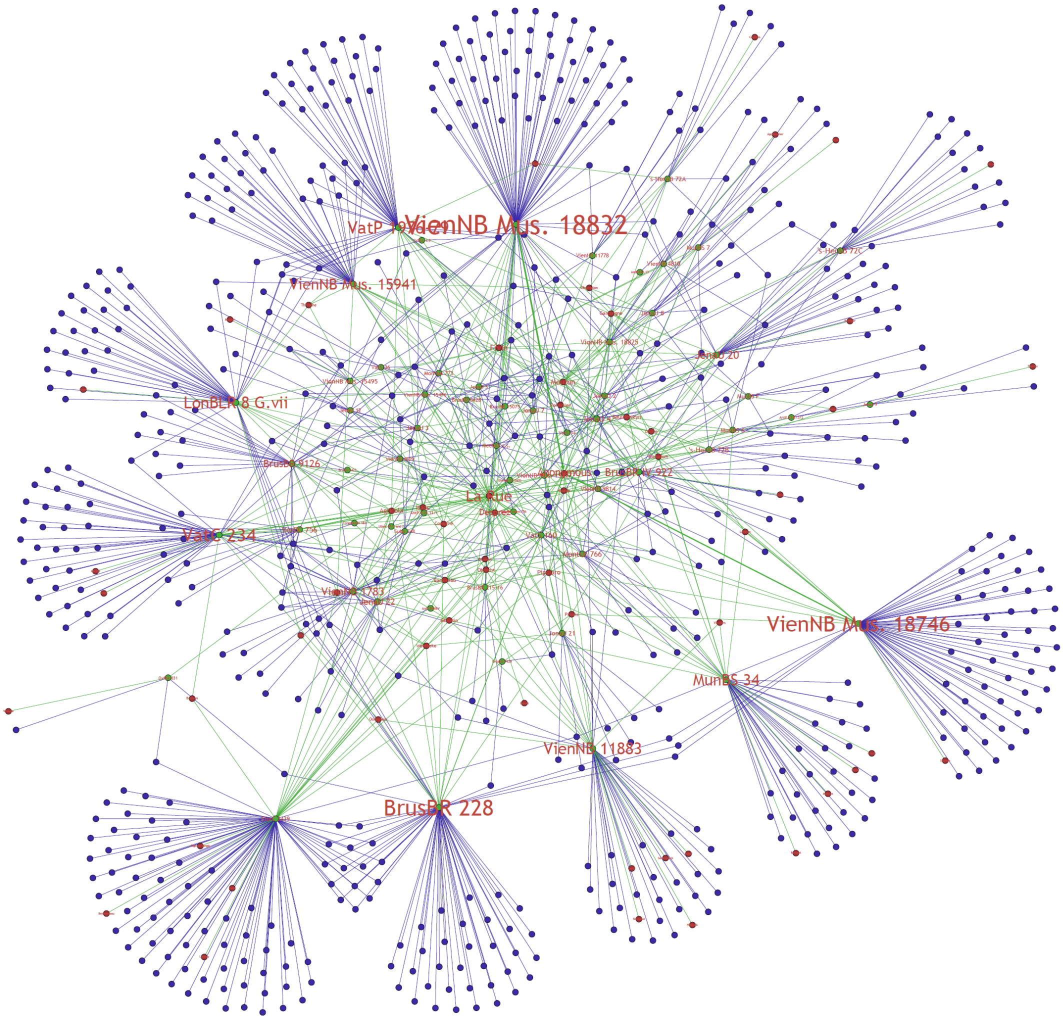


Figure 7.2: The network of the Burgundian-Habsburg music manuscripts; copied from the book of abstracts of the AHCN of 2013

7.2 Force-directed Approaches

Linked Sources: A Network Approach to The Repertory of Sixteenth-Century Polyphony [vB14]

At the *Arts, Humanities and Complex Network* conference in 2013, Marnix van Berchum presented a network approach to visualize Data taken from the court complex of music manuscripts of the Burgundian-Habsburg court. This corpus contains 51 complete manuscripts, 10 fragments and 600 compositions.

Interactive Visual Profiling of Musicians (Part B) [Jan15a]

Data. see part A 7.1.

Graph. The *Relationship Graph* is a social network graph including *all* relationships (the categories are: family of origin, partnership, education, relatives, godparenthood, affinity, personal relationships, working environment, dedication) of the selected musicians. Nodes represent the musicians, edges represent relationships between musicians.

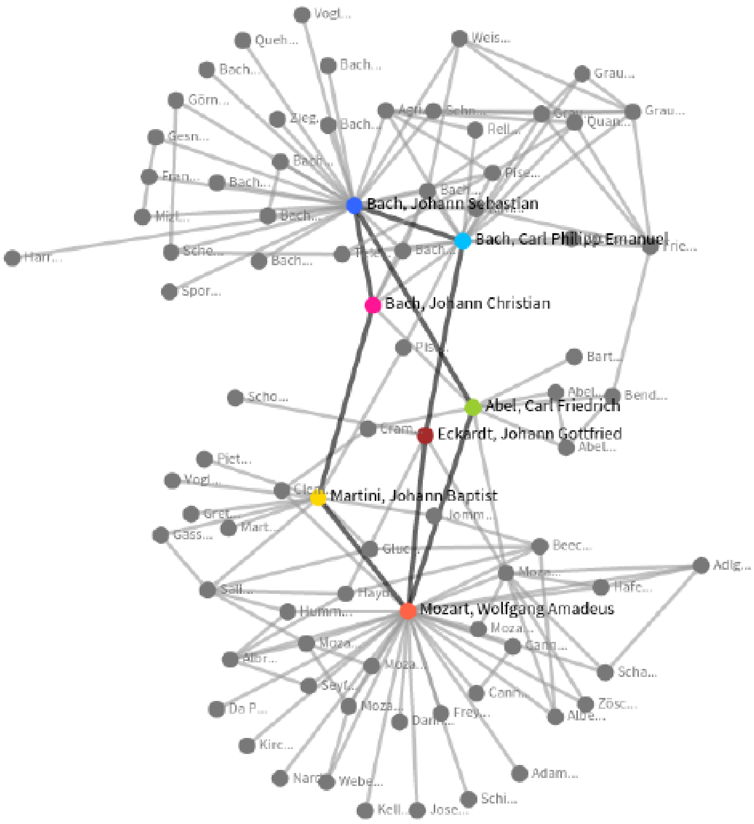


Figure 7.3: A Relationship Graph that links Johann Sebastian Bach to Wolfgang Amadeus Mozart; copied from [Jan15a]

Visualization. A force-directed algorithm was used for this relationship graph. The nodes of musicians are shown in their assigned colours, musicians that are added because they have a relationship to the selected musicians are shown in grey. The full names are only shown for the selected musicians. Additional information, like shortest paths to the other musicians, is displayed if a node is clicked on. In Fig. 7.3, an example for a relationship graph is shown. In this case, it displays potential connections between Johann Sebastian Bach and Wolfgang Amadeus Mozart.

8. Further Projects

Quantifying Kissinger

Quantifying Kissinger [1] is an online project by Micki Kauffmann which presents visualization of data regarding Henry Kissinger taken from the NSA's archives. Fig. 8.1 shows an excerpt from the interactive textplot of *telcons* - memoranda of telephone conversations - that was published on the blog [2]. The focus is on the word *office*.

The Senereko Project

The Senereko project [3] was a joint project of the Ruhr University in Bonn and the Digitial Humanities Institute in Trier. It used network analysis and visualization methods to survey the history of Religions based on three text corpora: The Pali Canon (a collection of buddhist texts), the Ancient Egyptian Thesaurus and the Mahabharata (an Indian Sanskrit epic).
Fig. 8.2 shows a network showing the central concepts of the first chapter of the Mahabharata. They are linked by community centrality.

[1] `http://blog.quantifyingkissinger.com/`
[2] `http://www.quantifyingkissinger.com/KA-site/`
[3] `http://senereko.ceres.rub.de/en/`

Figure 8.1: The Kissinger interactive textplot of telcons.

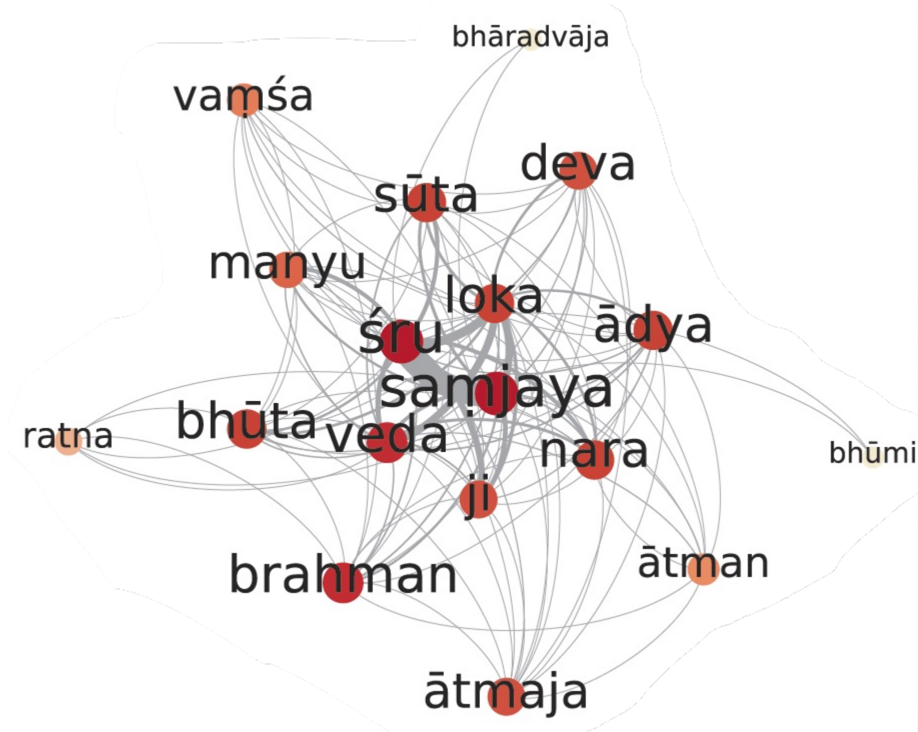


Figure 8.2: Visualization example, taken from a poster about the SeNeReKo project.

9. Conclusion

The field of digital humanities is a rather new and growing field of research. Yet, the inclusion of network analysis or data visualization techniques in several fields of the humanities, to visualize study findings has had a long history.
While we prepared the material of this thesis we observed that network visualization is not only useful for demonstrating findings of finished projects but also for the actual research. One good example is the poemage software which was developed in close cooperation with experts from the field of linguistics and literature. After they had been sceptical initially as regards the relevance of the use of network visualization for close reading analyses of poems, the authors state that the experts were even able to gain new insights into the poems by taking a different point of view [McC15].
The majority of visualization examples were created using a force-directed approach. Usually one of the common graph visualization software programs was used for them - and the options of the software determined what approaches were used to visualize the data.
In terms of different types of networks, the most common network type was the so-called co-occurrence network. It was used for literary texts as well as for letters or other data from archives in the fields of history. These networks reveal social structures in fictional or non-fictional texts and social network analyses can be performed on such networks and are actually performed frequently.
The field of archaeology is more experienced in the use of social network analyses. They have their own history of models using network techniques. For this survey, we only found examples using geographic visualization approaches which do not use network centrality measures.
In general, most examples for network visualization techniques were found in the field of literature. In this field social network analyses and distant reading combined with network visualization is very common. The use of visualization in close reading is still controversial but there are examples that suggest that it may be useful in the future.

Bibliography

[ABKS10] Evmorfia Argyriou, Michael A. Bekos, Michael Kaufmann, and Antonios Symvonis. On metro-line crossing minimization. 2010.

[Ahn13] Ruth Ahnert. Protestant letter networks in the reign of mary i: A quantitative approach, 2013.

[BC01] Ralf Brockenauer and Sabine Cornelsen. *Drawing Clusters and Hierarchies.* 2001.

[Bro16] Koenraad Brosens. Maptap and cornelia slow digital art history and fromal art historical social network research, 2016.

[Bru10] Tom Brughmans. Connecting the dots: Towards archaeological network analysis, 2010.

[Bru13] Tom Brughmans. Thinking through networks: A review of formal network methods in archaeology, 2013.

[BW13] Andrew Bevan and Alan Wilson. Models of settlement hierarchy based on partial evidence, 2013.

[CG72] E. G. Coffman and R. L. Graham. Otimal scheduling for two-processor systems. 1972.

[CHDS13] Cohen, Hersh, C. Dubay, and K. Spackman. Using co-occurrence network structure to extract synonymous gene and protein names from medline abstracts. 2013.

[CZQ+09] Weiwei Cui, Hong Zhou, Huamin Qu, Pak Chung Wong, and Xiaoming Li. Geometry-based edge clustering for graph visualization, 2009.

[dBT99] Giuseppe di Battista and Roberto Tamassia. *Layered Drawings of Digraphs.* 1999.

[Ead84] Peter Eades. A heuristic for graph drawing. 1984.

[EK86] P. Eades and D. Kelly. Heuristics for reducing crossings in two-layered networks, 1986.

[Els10] David Elson. Extracting social networks from literary fiction, 2010.

[Fen97] Qingwen Feng. Algorithms for drawing clustered graphs, 1997.

[FR91] Thomas Fruchterman and Edward Reingold. Drawing graphs by force-directed placement. 1991.

[GKNV93] E. R. Gansner, E. Koutsofios, S. C. North, and K.-P. Vo. A technique for drawing directed graphs, 1993.

[Gra14] Martin Grandjean. Intellectual cooperation: multi-level network analysis of an international organization, 2014.

[Gra15] Martin Grandjean. Introduction a la visualisation de donnees : l'analyse de reseau en histoire, 2015.

[Gra16a] Martin Grandjean. Archives distant reading: Mapping the activity of the league of nations' intellectual cooperation, 2016.

[Gra16b] Siobhan Grayson. Discovering structure in social networks of 19th century fiction, 2016.

[Gra16c] Siobhan Grayson. The sense and sensibility of different sliding windows in constructing co-occurrence networks from literature, 2016.

[GRPX16] Martin Grandjean, Yannick Rochat, Isaac Pante, and Aris Xanthos. Visualising the dynamics of character networks, 2016.

[HH99] Hadany and Harel. A multi-scale algorithm for drawing graphs nicely. 1999.

[Hil15] Vera Hildenbrandt. „im zentrum eines netzes [...] geistiger fäden". erschließung und erforschung thematischer zusammenhänge in heterogenen briefkorpora, 2015.

[HvW09] Danny Holten and Jarke J. van Wijk. Force-directed edge bundling for graph visualization, 2009.

[Jä15] Stefan Jänicke. Close and distant reading visualizations for the comparative analysis of digital humanities data, 2015.

[Jan15a] Stefan Janicke. Interactive visual profiling of musicians, 2015.

[Jan15b] Stefan Janicke. On close and distant reading in digital humanities: A survey and future challenges, 2015.

[JGS16] Stefan Janicke, Anette Gessner, and Gerik Scheuermann. A distant reading visualization for variant graphs, 2016.

[KK89] Kamada and Kawai. An algorithm for drawing general undirected graphs, 1989.

[Kna08] Carl Knappett. Modelling maritime interaction in the aegean bronze age, 2008.

[Kob13] Kobourov. *Force-directed drawing algorithms.* 2013.

[LBA10] A. Lambert, R. Bourqui, and D. Auber. Winding roads: Routing edges into bundles, 2010.

[McC15] Nina McCurdy. Poemage: Visualizing the sonic topology of a poem, 2015.

[Mor05] Franco Moretti. *Graphs, Maps, Trees: Abstract Models for a Literary History.* 2005.

[Mor13] Franco Moretti. *Distant Reading.* 2013.

[Noe10] Martin Noellenburg. An improved algorithm for the metro-line crossing minimization problem, 2010.

[Pur02] Helen C. Purchase. Metrics for graph drawing aesthetics. 2002.

[Roc14] Yannick Rochat. A network analysis approach of the venetian incanto system, 2014.

[Roc15] Yannick Rochat. Character network analysis of Émile zola's les rougon-macquart, 2015.

[Sua12] Juan Luis Suarez. Towards a digital geography of hispanic baroque art, 2012.

[Sug81] Sugyama. Methods for visual understanding of hierarchical system structures. Feb. 1981.

[vB14] Marnix van Berchum. Linked sources: A network approach to the repertory of sixteenth-century polyphony, 2014.