Aalto University
School of Science
Degree Programme in Information Networks

Noora Routasuo

Exploratory Visualization of Inter-Organizational Networks

The Visualization Process

Master's Thesis
Espoo, June 17, 2013

Supervisor: Professor Tapio Takala, Aalto University
Advisors: Tomi Kauppinen, PhD
Annukka Jyrämä, PhD

Aalto University
School of Science
Degree Programme in Information Networks

ABSTRACT OF
MASTER'S THESIS

Author: Noora Routasuo

Title:
Exploratory Visualization of Inter-Organizational Networks The Visualization Process

Date:	June 17, 2013	**Pages:**	86
Major:	Media Technology	**Code:**	IL3011
Supervisor:	Professor Tapio Takala		
Advisors:	Tomi Kauppinen, PhD Annukka Jyrämä, PhD		

Information visualization is a powerful tool for communicating complex ideas, but also for exploring data. Research in information visualization has been fueled by the continued growth in the size and complexity of data sets, but it has focused mainly on visualization techniques. Understanding the process of visualization from a wider perspective would support both the development of visualization software, and the adaptation of information visualization as an exploratory technique. This thesis aims to study this process, and how it can be used to support the exploration of inter-organizational networks in particular.

The visualization process was examined in a literature review based on several related disciplines. In addition, a case study involving the visualization of inter-organizational networks was conducted. The case study consisted of the development of two visualizations and the examination of the exploratory visualization process through interviews with domain experts in an organizational context.

As a result of the literature review, a synthesis process model of exploratory visualization is presented and its steps are described in detail. This model is compared to the results of the case study. Based on the case study, design implications for effective visualization of inter-organizational networks are discussed.

The process model demonstrates that there is much more to exploratory visualization than creating visualizations. It includes data collection and processing, interpretation, evaluation, and communication of exploratory results. The process resembles that of data mining, but there are some key differences, such as the flexibility of the interpretation step, potential for collaboration, and the nature of the results. For the visualization of inter-organizational networks, important considerations are the definition of a relationship between two organizations, evolution over time, and showing data about individual organizations.

Keywords: information visualization, visual data mining, inter-organizational networks, cooperation networks, visualization process, exploratory visualization, data analysis

Language: English

Aalto-yliopisto
Perustieteiden korkeakoulu
Informaatioverkostojen koulutusohjelma

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Noora Routasuo

Työn nimi:
Organisaatioiden välisten verkostojen exploratiivinen visualisointi Visualisointiprosessi

Päiväys:	17. kesäkuuta 2013	**Sivumäärä:**	86
Pääaine:	Mediatekniikka	**Koodi:**	IL3011
Valvoja:	Professori Tapio Takala		
Ohjaajat:	Tomi Kauppinen, FT Annukka Jyrämä, KTT		

Informaation visualisointia voidaan käyttää työkaluna tiedon esittämiseen, mutta myös dataan perehtymiseen (*exploration*). Data-aineistojen kasvu ja monimutkaistuminen on toiminut kannustimena informaation visualisoinnin kehittymiselle, mutta tähänastinen tutkimus on keskittynyt visualisointitekniikoihin. Visualisointiprosessin ymmärtäminen kokonaisuutena tukisi sekä visualisointiohjelmistojen kehittämistä että visualisoinnin käyttöönottoa dataan perehtymisen työvälineenä. Tämän diplomityön tavoitteena on tutkia tätä prosessia, ja sitä, kuinka se voi tukea organisaatioiden välisten verkostojen hahmottamista.

Visualisointiprosessia tarkasteltiin poikkitieteellisessä kirjallisuuskatsauksessa. Lisäksi toteutettiin organisaatioiden välisiä verkostoja koskeva tapaustutkimus. Tutkimuksessa kehitettiin kaksi visualisaatiotyökalua ja tutkittiin exploratiivista visualisointiprosessia kohdeorganisaation asiantuntijoiden kanssa järjestetyissä haastatteluissa.

Kirjallisuuskatsauksen pohjalta esitetään synteesinä exploratiivisen visualisoinnin prosessimalli ja kuvataan se vaihe vaiheelta. Tätä teoreettista mallia verrataan tapaustutkimuksen tuloksiin. Tapaustutkimuksen pohjalta esitetään myös suuntaviivoja organisaatioiden välisten verkostojen tarkoituksenmukaiseen visualisointiin.

Prosessimalli osoittaa, että exploratiiviseen visualisointiin liittyy paljon muutakin kuin pelkkä visualisointien luominen. Siihen kuuluu datan kerääminen ja esikäsittely, visualisaation tulkinta ja tulosten viestintä. Prosessi muistuttaa tiedonlouhintaa, mutta näissä on myös eroja, kuten tulkintavaiheen moniulotteisuus, yhteistyön mahdollisuudet ja tulosten luonne. Organisaatioiden välisten verkostojen visualisoinnissa on tärkeää organisaatioiden välisen suhteen määrittely, verkoston kehitys ajan mittaan ja yksityiskohtien näyttäminen yksittäisistä organisaatioista.

Asiasanat:	informaation visualisointi, visuaalinen tiedonlouhinta, organisaatioiden väliset verkostot, yhteistyöverkostot, visualisointiprosessi, exploratiivinen visualisoiti, data-analyysi
Kieli:	englanti

Acknowledgements

This thesis was carried out at the Department of Media Technology in Aalto University School of Science.

I would like to thank my instructor Tomi Kauppinen for enthusiasm and inspiration thorough the process of writing this thesis, and my supervisor Tapio Takala for encouragement and ideas. This work would not have been possible without their support. I am also indebted to my second advisor Annukka Jyrämä and Atso Andersen from the Institutional Relations unit of Aalto University for their contributions to the project, especially for their participation in the case study. Finally I would like to thank the developers of the various visualization tools that I have been able to use in my work.

Espoo, June 17, 2013

Noora Routasuo

Contents

Part I

Preliminaries

Chapter 1

Introduction

Information visualization can serve as a tool for three distinct purposes: communication, analysis, and exploration. When used as a communicative tool visualization aims to clearly and precisely communicate complex ideas to the viewer. In analysis, information visualization is used to confirm or test hypotheses, to compare and contrast, and to gain insight to a particular problem. Finally, in exploration, visualization is used to create ideas that can lead to hypotheses about a data set that is not yet well understood.

Exploratory visualization is becoming increasingly relevant as the amount and complexity of raw data available to organizations continues to increase exponentially. The bulk of the research on information visualization focuses on developing techniques and systems for visualizing this data. Less attention is given to the overall process of exploratory visualization and its practical implications. This thesis examines the visualization process itself through a literature review and a case study involving the visualization of Aalto University's cooperation network.

1.1 Motivation

Visualization is a powerful method for exploring data because it combines the superior visual abilities of humans with the computational power of machines. The process of exploratory visualization starts with the collection of data and ends with new ideas and hypotheses that form a basis for further analysis and debate. Understanding exactly what happens between would not only speed up the process, but lead to the development of better tools to support it [Kandel et al. 2012, Neumann et al. 2007]. Exploratory visualization, if understood on this level, could then benefit both science (Kandel et al. [2012] goes as far as to say that visualization is a major bottleneck in scientific

progress) as well as business (where visualization typically supports two main tasks: monitoring and decision-making [van der Heijden 2009]).

Especially for businesses, but also other organizations such as universities, understanding the inter-organizational networks that they operate in is of particular interest. This are an example of a topic whose analysis requires vast amounts of data that most likely already exists in some form in said organizations, but has not been systematically addressed. While *network theory* provides a theoretical background for the analysis of such networks, exploratory visualization can provide analysts with an overview of the data they are working with, and managers and decision makers with the practical tools to understand their position and influence. To be practical, however, such tools cannot be just excellent visualization algorithms – they must take into account the whole process of exploratory visualization in an organizational context.

1.2 Research questions

The aim of this thesis is to understand the visualization of a particular kind of data – data about inter-organizational networks – at an early stage of the analysis, and to investigate how exploratory visualization can be used to understand such networks and support decision-making. The research questions can be formulated as follows:

1. What steps does the exploratory visualization process include and what are they like?
2. How can this process support the analysis of inter-organizational networks?
   (a) What kind of visualizations are useful in this context?
   (b) What kind of ideas and hypotheses can exploratory visualization produce about such networks?

Answers are sought from two directions. First, the overall process of exploratory visualization is outlined through a literature review which will result in a synthesized process model. Second, the process is further examined through an exploratory case study about the visualization of the inter-organizational cooperation network of Aalto University. The case study involves iterative development of visualizations and interviews with domain experts on the things that those visualizations reveal or do not reveal about the network, thus suggesting requirements for different steps of the process (such

as data collection) and relevant themes in the analysis of inter-organizational networks. These two perspectives – general and specific – both contribute to the first question. The second question is primarily considered in the case study, but an overall understanding of the process will give us a context in which to formulate the answer.

This study does not aim to be a complete review of visualization techniques or tools. Our focus is on exploration, not technical finesse or performance. Although new visualizations are developed as a part of the case study, developing new visualizations is not the main goal, but rather we are interested in the general kind of visualization approach that fits a specific purpose. Detailed network analysis will also be left for future research as our context here is an early stage of the analysis of unexplored data.

The main contribution of this work is a proposed process model of visual exploration, based on relevant models and characterizations in previous literature, and a qualitative case study involving exploratory visualization in a real organizational setting. These reveal several tasks and challenges in exploration not typically discussed in the context of visualization, and provide guidelines for visualization system designers, analysts, and managers.

1.3 Structure of the thesis

This thesis is structured in four parts as follows. Chapter 2 gives an overview of data analysis, information visualization and the management challenges related to inter-organizational networks (Part I). Chapter 3 details the methodology and context of the case study, and Chapter 4 reviews existing literature on the process of exploratory visualization data mining (Part II). A synthesis process model based on the literature review is presented in Chapter 5 and the results of the case study in Chapter 6 (Part III). Finally, findings are combined and discussed in Chapter 7 and conclusions are drawn in Chapter 8 (Part IV).

Chapter 2

Background

To understand the context of this research, we first review key concepts in data analysis, information visualization, and inter-organizational networks. Section 2.1 introduces the challenges in data analysis that create the need for analytical tools such as information visualization. Common tools and techniques in visualization as well as the characteristics of exploratory visualization are discussed in Section 2.2. Networks and particularly inter-organizational networks are defined in the context of this study and a brief overview of complex network theory is given in Section 2.3.

2.1 Data and discovery

In the past two decades, two trends related to data have risen that are shaping the way institutions and individuals operate. On one hand, data has exponentially increased in volume, both in globally and in terms of data produced and consumed by single organizations as a part of their normal operation, to the extend that only small amounts of it will ever be used [Myatt and Johnson 2009, Kantardzic 2011]. On the other hand, data has also become increasingly available in various formats due to advances in technology, particularly the ever-increasing storage space [Hilbert and López 2011].

All this data is collected and stored because business analysts and scientists alike believe that valuable information is hidden in it, but extracting this information is not a trivial matter. Data sets are often simply too big or too complex to be analyzed manually – thousands or hundreds of thousands of instances with hundreds of inconsistent variables – so various methods for data analysis are becoming indispensable.

2.1.1 Data mining

Already in 1983, referring to data in organizations, Wildavsky [1983] wrote: "Just as mountains are climbed because they are there, more data are produced because it is possible." He was concerned that modern information systems undermine the ability of organizations to transform data into genuinely meaningful bits of information. Since then, the amount of data has only grown, but effective means of analysis have also been developed.

Data mining is a discipline firmly founded in statistics aiming to extract new patterns or models from available data. Originally focused on databases, filtering and statistical modelling, data mining has come to include a number of new techniques arising from machine learning, natural language processing, and visualization [Kantardzic 2011]. Data mining that makes extensive use of visualization is often referred to as *visual data mining* [Keim 2002].

Data mining covers both exploration and more detailed analysis of data. It has two main goals: prediction and description [Myatt and Johnson 2009]. Description involves gaining insight in the current state of things such as finding associations and detecting unusual patterns, while prediction involves classification and regression models. Kantardzic [2011] defines data mining as a "cooperative effort of humans and computers" and lists the primary tasks of data mining as: classification, regression, clustering, summarization, dependency modeling and change and deviation detection. All of these tasks involve specific techniques borrowed from statistics and machine learning, and produce distilled facts from the data.

Besides the challenge of the size and complexity of data, analysts face other problems as well. Data may be incomplete or haphazardly collected. Fields may be used inconsistently when recording data and entities may be referred to with different names. For example, the name of a person may be written in several different ways. Data may also be scattered – for example, data collected by different parts of an organization but inherently describing the same thing can be formatted in subtly but significantly different ways [Han et al. 2006]. Dealing with these issues in one way or another – either discarding unideal data or tediously converting it to an applicable format – is a part of the data mining and analysis process.

2.1.2 Linked Data

While data mining is often focused on databases and data warehouses, *Linked Data* is an example of another type of data that is increasingly available. It represents an entirely different paradigm of sharing and publishing data to for anyone to access, analyze, and visualize.

According to Heath and Bizer [2011], Linked Data aims to utilize the World Wide Web to create a "Web of Data" where "not only documents, but also data, can be a first class citizen of the Web". Web pages are designed to be human-readable and require extensive pre-processing to be useful for automated data analysis. However, the Web is a platform where data can be published, accessed and found easily anywhere, and can be linked to connect separately published parts into a large, meaningful whole. Linked Data is a model for publishing data that is driven by open standards to achieve the consistency required for such a global data space.

Linked Data is based on the Semantic Web, collection of technologies to publish semantic content on the Web. [Heath and Bizer 2011] It includes a number of technologies in different stages of development and adoption. The most important technologies for Linked Data are standards for describing entities, links, and vocabularies, as well as various serialization formats.

Since Linked Data can be published by anyone, in the ideal case, data sets will grow to complement each other, creating a large database available for anyone to use. In reality, there are issues of credibility as well as issues of maintenance similar to databases [Bechhofer et al. 2013]. Nonetheless, it is a growing paradigm of data publishing that facilitates sharing and reuse and drives the development of data analysis, and is particularly relevant for the study of networks due to its linked nature.

2.1.3 Information systems and data aggregation

Wildavsky [1983] defines organizations as systems that suppress data. They filter and aggregate it and produce summaries for the management. Similarly any data mining or analysis, whether for the purposes of management, science or something else, produces some kind of an aggregated result, which is often presented in the form of a report or summary [Myatt and Johnson 2009]. Various information systems support the exploration of data in this manner. These come in many flavours but we consider the management perspective here because it is particularly relevant to inter-organizational networks and our case study. Much of what applies to management information systems, however, is applicable for any information system.

Management information systems (MIS) are systems that produce the information that managers use to make important strategic decisions. At their simplest they are spreadsheets perhaps powered with a few illustrating charts while complex management information systems incorporate various reports and data mining elements. [van der Heijden 2009] According to van der Heijden [2009], typical tasks that management information systems support managers with are monitoring key performance indicators (such as

total revenue or sales) and selecting alternatives (when making decisions).

It is worth noting how similar this definition is to the two goals of data mining – description and prediction – mentioned earlier. In fact, the aim of these systems is fundamentally the same as that of data mining. The amount of data a human (manager or not) can accurately process is limited by our cognitive abilities [Spence 2007]. Effective decision-making requires distilled information. Thus, the single most important task of a management information system is the aggregation of relevant data [van der Heijden 2009]. This aggregation is the heart of analyzing complex data sets.

2.2 Information visualization

Information visualization is the transformation of abstract data to a presentation that is more readily understood by humans. It is a powerful approach to making sense of complex data, both for communicative and analytical purposes. While information visualization has only fairly recently emerged as an independent field of study, it has long roots in innovative use of visual presentations for communication and analysis thorough history [Tufte 1983; 1990, Freeman 2000]. Recently, computation has vastly improved the efficiency and analytical capability of visualizations, so that they can be created with less effort and represent more data and dimensions than before, and can include real-time interaction [Spence 2007]. This section discusses the uses, tools and evaluation of visualization in general, and the characteristics of exploratory visualization in particular.

2.2.1 The power of visualization

While according to Spence [2007], information visualization need not only refer to *visual* but could also encompass a variety of presentation techniques, the field is most often concerned with leveraging the superior visual capabilities of humans. Some see it as a potentially ideal combination of human analytical skills and flexibility on one hand, and the computational power of machines on the other [Shneiderman 1996, Keim 2002].

Visualization is arguably one of the most important tools of science [Freeman 2000]. A good visualization can convey vast amounts of information, easing analysis, decision-making and the conveying of complex ideas. As Keim [2002] points out, visualization can especially better deal with heterogeneous and noisy data than pure data mining. Bad visualizations on the other hand can, intentionally or not, be extremely misleading [Tufte

1983]. The recent research in information visualization has focused on the computer-assisted creation of visualizations of various data.

Visualization is a powerful tool not only for communicating existing ideas, but also for exploring large data sets. This notion has been on the forefront of research into information visualization for the past decades. The development of computation and the increase of data volumes has led to the intertwining of information visualization with data mining [Keim 2002]. Information visualization is now a common part of the workflow of specialist data analysts [Kandel et al. 2012] and information visualization software aimed at specialists are a multitude. In fact this paradigm is so strong in research that the term *visualization* is often used to refer to (the development of) *visualization systems* for scientific and business purposes.

On the other hand, the same context – computation and the availability of data – have created a completely opposing trend that some authors [Pousman et al. 2007] call *casual information visualization*. This refers to the appearance of visualizations in magazines and newspapers, as well as the simple visualization tools aimed at casual users who want to, for example, see a visual representation of their own social network. Even quite complex exploratory visualizations systems aimed at regular users "just for fun" have been successfully developed (for example, by Heer and Boyd [2005]).

However, although commercial tools do exist, complex visualizations have not yet been widely adopted outside of the scientific community [Amar and Stasko 2004, Plaisant 2004]. Several authors have argued that this is because the development of information visualization systems is focused on presentation without truly being able to provide practical insight to the users. Amar and Stasko [2004] believe this is due to lack of advanced analytical features in such systems, leading to "analytic gaps" between the presentation and a user's ability to derive insight from them. Jankun-Kelly et al. [2007] on the other hand argue that systems lack support for the whole analysis process.

2.2.2 Tools and techniques

The growing demand and diverging needs of visualization are reflected in the multitude of techniques and tools available. The choice of tools ultimately depends both on the goals of the visualization, but also with the preferences and skills of the visualizer.

Visualization techniques range from simple scatter plots to complex, multiview visualization systems, from immediately intuitive displays, such as time series and mosaic plots, to mathematically-grounded analytical visualization techniques capable of fitting vast amounts of data into a single view but requiring familiarizing, such as *self-organizing maps* (SOM). The choice of the

technique is dependent on the data (type, amount, quality) to be visualized and the purpose of the visualization [Shneiderman 1996, Keim 2002]. For a thorough review of various visualization techniques, the reader is referred to a visualization textbook such as [Spence 2007] or the thorough review by Keim [2002].

Software tools are as varied as techniques. Basic visualizations can be produced by a variety of out-of-the-box or web-based software such as Google Charts [Google] or Circos [Krzywinski et al. 2009], and simple analytics can be performed by a number of free applications such Gephi [Bastian et al. 2009], SoNIA [McFarland and Bender-deMoll] and many, many others. These vary in their complexity, ability to handle large data sets, and flexibility, interactivity of the visualization, and data types and techniques supported. Commercial and academic software provide increasingly complex analytical capabilities. Academically developed visualization software typically feature extended data handling, multiple coordinated views and filtering, with the expense of a more cluttered user interface [Kang et al. 2007, Callahan et al. 2006]. Basic visualization tools often also often also come embedded in other analytical software such as spreadsheet software [van der Heijden 2009].

For maximum flexibility in developing visualizations, programmers can rely on statistical environments such as Matlab [MATLAB 2010] or R [R Core Team 2013] that offer many visualization packages, or programming frameworks for any language (such as Java, JavaScript, Python, C, or Flash) that fits their purposes.

2.2.3 Exploration

Three different aims for visualization can be distinguished. *Presentation* is the communication of distinct ideas to the viewer. It is the strength of visualization in that other data aggregation methods such as data mining do not inherently support presentation as such. Second purpose is *analysis and confirmation*. Given a data set, these visualizations try to answer a specific question such as if a particular phenomenon has grown in a particular time frame. John Snow's analysis of a cholera epidemic in London in 1854 as narrated by Tufte [1997, pp. 27–37] is a classical example of hypothesis-testing and analytical visualization. Finally, the third purpose of visualization is *exploration*.

The aim of exploration is the generation of ideas and hypotheses rather than communicating or testing a specific idea. It often means the exploration of a specific data set and is closely related to data mining and analysis. In literature, exploratory visualization often involves special-use software with a seamless flow from one visual presentation to another, and is conducted by

experts such as chemists, biologists, or data analysts [Demoll and Mcfarland 2005, Neirynck and Borner 2007, Callahan et al. 2006]. These visualization systems do not necessarily even create a visual end result, but rather, insight is created during interaction with the system itself. The visualization techniques used are limited to those offered by the system, but multiple data sets can be explored using the same system.

However, it is useful to distinguish two more types of visual exploration. First, at its simplest, exploration can refer to the way a user makes discoveries while exploring a complete interactive visualization. For example, considerable insight can sometimes be gained by simple interaction techniques such as zooming, filtering and selecting [Heer and Boyd 2005]. In this case, the data set and visualization technique have already been chosen and the user can only tweak some parameters of the presentation. A third type of exploration involves the iterative creation of independent visualizations to be viewed and discussed collaborative and then further refined to bring out more detail about an interesting theme. This type of exploration will be the main technique used in our case study, detailed in Chapter 3. This informal typology of exploratory visualization is summarized in Table 2.1.

Table 2.1: Types of visual exploration.

	Data fixed	Technique(s) fixed
Scientific	no	yes
Interactive	yes	yes
Iterative	no	no

These three approaches to exploration can naturally overlap, for example when iteratively created visualizations are also interactive, or when a specialist working with a visualization software pauses to show an intermediate result to someone else who might have suggestions as to the direction of the exploration. It is useful to make the distinction however, as we consider the visualization process and try to pinpoint when and how insight is created.

2.2.4 Evaluation

If the purpose of visualizations is to deliver ideas and facts and to support analysis, then designers of visualizations have a responsibility to ensure the correctness of their presentations. Tufte [Tufte 1983; 1997] provides many examples of the misleading and destructive potential of purposely skewed or unintentionally confused visualizations. He argues that imprecision and

lack of documentation is more often forgiven in graphics where it would be unacceptable in text and tables. One reason that it might be so is that we are better trained to be critical of words and tables, and the high level of aggregation in visualization can potentially obscure the reasoning behind it [Amar and Stasko 2004]. Another is that evaluation of visualizations often lacking even in the research of information visualization.

Several researchers [Chen and Czerwinski 2000, Ellis and Dix 2006] describe an urgent need for new methodologies for the evaluation of visualizations. In a literature sample, Ellis and Dix [2006] found that the majority of papers describing new visualization techniques do not mention evaluation at all, or place it under "Future work". Yet evaluation is relevant to the field of information visualization as a whole, with the constant flow of new techniques and tools, as well as to the user who must be convinced that a given visualization is accurate, useful, and worth their time and money [Carpendale 2008].

Existing evaluation methodologies of information visualization fall into three categories [Sedlmair et al. 2011, Plaisant 2004]:

Usability evaluation adopted from HCI (human-computer interaction) and focused on the user interfaces of visualization tools

Comparison of design elements in terms of human perception or usability

Comparison of one or more tools with some state-of-the-art

Case studies in realistic settings, including longitudinal studies

Usability and user interface issues are relevant for visualizations, especially complex visualization software, because they are often interactive. They are particularly useful in finding things to develop in a particular tool or visualization, but their evaluation is usually based on artificial tasks which limits their generalizability – [Plaisant 2004] argue that longitudinal case studies are more realistic than evaluation based on artificial tasks, since complex tools are usually meant for long-term use. Longitudinal studies are, however, difficult to arrange and thus rare.

Perception-based studies are closest to answering the question of the *correctness* of a visualization. They can involve, for example, the comparison of two visual elements to see if users correctly perceive their relative size. [Heer and Bostock 2010] report an interesting experiment in crowd-sourcing these kind of studies.

A different approach to evaluation visualizations is suggested Tufte [1983, pp. 91–105] with the concept of "data-ink ratio", which is the ratio of ink (pixels) dedicated to conveying actual information relative to the total ink (pixels) used by the visualization. A cluttered graph with non-necessary decorations would have a low data-ink ratio while a succinct box plot would have a ratio close to one.

The evaluation of exploratory visualization, however, poses different questions to those covered by the evaluation techniques discussed thus far. Besides perceptually correct, exploratory visualizations are supposed to provide insight and ideas. The task is not well-defined, and the results are hard to measure. Plaisant [2004] suggests that case studies are the best approach to evaluating exploratory visualization (and lists come benefits and disadvantages of case studies for evaluation in general).

The evaluation of visualizations is complex, because ideas communicated by them are dependent on the users' abilities and assumptions, the data used, the system as a whole, and other factors [Ellis and Dix 2006]. Much of the correctness of a visualization draws from the correctness of the underlying data. If the data is skewed, so is the visualization.

2.3 Network theory and inter-organizational networks

Network is an ambiguous term that, in the context of computer science, usually refers to one of two things – either computer networks or a type of data that represents relationships between some entities. While the former can be thought of a subset of the latter, in research these two often represent very different perspectives. In the context of this study we are focusing specifically on data about relationships, particularly inter-organizational networks.

In the following, we first review features of network theory that is relevant to the analysis of inter-organizational networks and then discuss inter-organizational networks themselves. While in the exploratory context of this work we will not go to an in-depth analysis, this will allow us to understand the underlying ideas that are of interest in this particular type of data.

2.3.1 Graphs

Networks are often modeled as *graphs*, an abstract representation which is founded in mathematics but has also become a standard tool of network visualization. Graphs are defined as consisting of *nodes* (also called *vertices*)

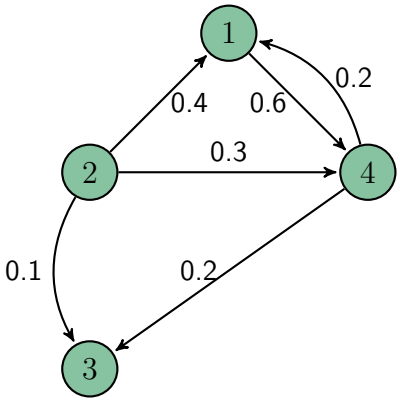

Figure 2.1: Example of a simple (directed and weighted) graph.

and *edges* (also called *links*) between those nodes. The *degree* of a node is the number of edges connected (incident) to it. Edges can be weighted or directed – usually the whole graph is considered either directed or undirected [Newman 2003]. *Weighted* edges and represent things like the strength of a relationship or the amount of traffic going through the edge. *Directed* edges represent things like citations or traffic, while *undirected* edges represent mutual relationships such as friendship [Easley and Kleinberg 2010]. Figure 2.1 shows a small directed and weighted graph with four nodes and six edges.

In a given graph, edges may or may not be allowed to form cycles. The example graph in Figure 2.1 does contain a cycle, since it is possible to go from node 1 to node 4 and back again. A graph with no cycles is called a *tree* and inherently represents a hierarchy. A group of separate trees is called a *forest*. Nodes may also be of different types and have attributes. For example, if a node represents a person, attributes can include age, gender, occupation and so forth.

2.3.2 Characteristics of complex networks

A *complex network* is loosely defined as a network with non-trivial topological features. Such networks arise naturally in many contexts such as sociology, biology and mathematics, and they have some common properties that are quite different from artificial networks such as computer networks on one hand and entirely random networks on the other [Easley and Kleinberg 2010]. The study of these properties and complex networks in general is a fairly new and rapidly growing field. Recently, particularly social networks have become an increasingly active field of study, and the focus of research has shifted

from small networks (less than 100 nodes) and the properties of individual elements to large networks and their statistical properties [Newman 2003].

According to Caldarelli and Vespignani [2007, pp. 15–16], the complexity of complex networks arises from the fact that they are self-organizing and exhibit emergent architecture and complex topological features. Complex networks do they have structure (unlike random networks) but the structure is not entirely regular (like artificial networks tend to be).

An example of this structure is the *small world property*, which means that the shortest path from one node to another is small compared to the size of the network. The concept has been popularized in social networks as "six degrees of separation", referring to the supposed six links that connect every person on the planet. Complex networks also exhibit *clusters*, groups of nodes that are highly inter-connected, such as cliques of friends. Finally, the degree distribution of the nodes of complex networks tends to follow the power law, that is, there are many nodes with few connections but it is also possible to find few central nodes with an extremely high degree. [Caldarelli and Vespignani 2007, pp. 11–15]

As a fundamental data that represents the kind of information that is extremely relevant to many organizations – relationships – networks offer many interesting perspectives. Ultimately, the study of networks aims to understand the underlying systems that they represent. For example, the study of resilience – how networks survive when nodes are removed – aims to improve the stability or security of infrastructure [Newman 2003].

2.3.3 Visualization of networks

Shneiderman [1996] identifies seven different data types for the purpose of visualization: 1-dimensional, 2-dimensional, 3-dimensional, temporal, multi-dimensional, tree, and network. While such taxonomies are simplifications – real-life data can be a combination of types, and visualizations can be applicable to several types or only a special case of one – clearly data type affects the choice of a visualization approach.

By far the most common technique for visualizing networks is graphs, also referred to as node-link diagrams [Kang et al. 2007, p. 215]. In the node-link diagram, network actors such as organizations or persons are depicted as nodes, while their relationships are edges (lines) between nodes. Lines and nodes may be coloured or labeled to convey additional information. Key issues in creating graph-based visualizations is finding a balanced layout of the nodes and avoiding overlapping of nodes and edges as much as possible. The importance of these considerations grows with the size of the network to be shown. Various layout algorithms aim to satisfy additional requirements

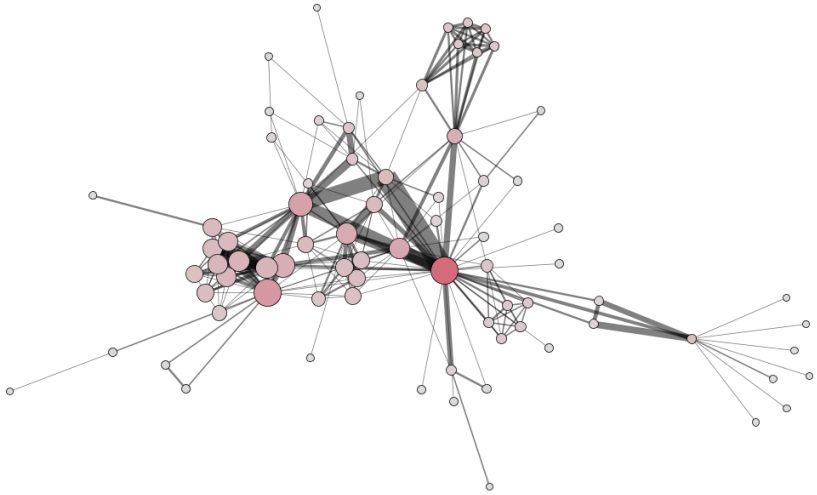

Figure 2.2: Visualization of a larger graph using Gephi.

such as placing related nodes close to one another, maximizing symmetry, equal or weight-dependent edge lengths, maximum node orthogonality, and so on [Šperka and Kapec 2010]. Figure 2.2 presents and example of a graph visualization created with the Gephi [Bastian et al. 2009] software.

Another common technique, also based in mathematics, is matrices. A network of n nodes is presented as a $n \times n$ matrix where element a_{ij} represents the relationship between nodes i and j. Matrices are extensively compared with node-link diagrams by Ghoniem et al. [2004], who conclude that the techniques are complementary and that matrices are better suited for large or dense networks. Kang et al. [2007] note that while node-link diagrams are useful in topology-based tasks such as finding clusters or detecting shortest paths, they are poor choice for attribute-based tasks.

Entirely different approaches to network visualization are also possible. Kang et al. [2007] present a visualization system largely built around multiple coordinated bar charts and tables (Figure 2.3). The visualizations themselves are very simple, but their coordination allows the user to grasp the complex relationships. The system was developed to represent co-authorship networks and is conceptually based on an "actor-content" data model – that is, it allows for a specific kind of visualization by making an assumption about the structure of the underlying network. It is an example of an exploratory visualization system that does not produce a definitive visual end-result but allows the user to understand a data set by exploring it within the system. Other approaches include visualizing networks as trees [Card et al. 2006, Lee

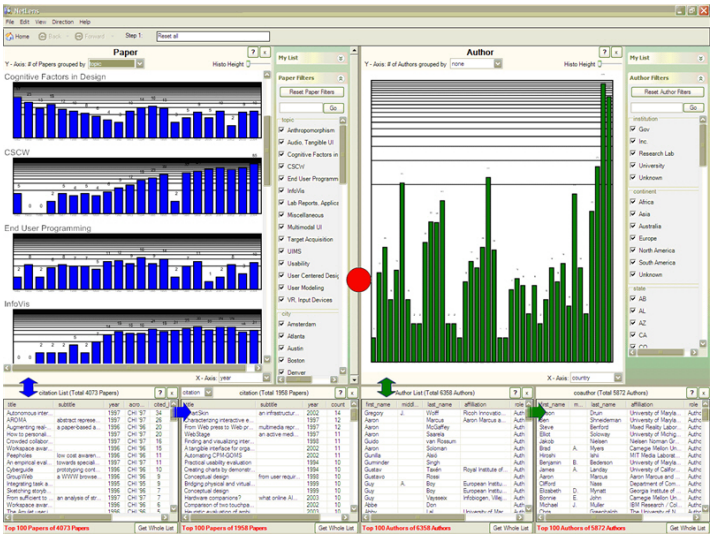

Figure 2.3: NetLens, a network visualization system by Kang et al. [2007]. [2]

et al. 2006] by only showing relevant portions at a given time, and *hive plots* [Krzywinski et al. 2012] that aim to make the positioning of the nodes more meaningful.

Freeman [2000] gives a fascinating account of the history and progress of the visualization of social networks. From it we can trace the progression as in visualization overall - from hand-drawn presentations whose power and clarity rely solely on the expertise and finesse of the creator, to the emergence of computation, growing data sets, interaction and, finally, the web and the new trend of casual visualization parallel to that of scientific visualization.

2.3.4 Social and organizational networks

Social networks are a type of complex network that have been a focus of much research of complex network theory since its very beginnings. In social networks, nodes are usually considered to be individuals (although they can also be connecting entities such as hobbies and places or hobbies [Easley and Kleinberg 2010]) and edges their relationships. They are accessible and important real-world networks that arguably influence every singly field of human activity.

An interesting feature of social networks is that, unlike many other naturally occurring networks (such as biological networks), they tend to be as-

[2]Screenshot from `http://www.cs.umd.edu/hcil/netlens/`

sortative [Newman 2003] – that is, nodes have a tendency to be connected to others that are somehow similar to them. This similarity could be based on ethnicity, gender, hobbies or any other attribute and has far-reaching implications for how the networks function. Another interesting feature is the strength of a ties, since close or more casual ties between individuals imply very different relationships. Due to assortativity, weak ties are often important links between communities that are crucial in the spread of information, disease, or any other matter under study [Easley and Kleinberg 2010].

Collaboration and affiliation are much-studied phenomenon in social networks. An example of a collaboration network is the co-authorship network among academics, where authors who have written a paper together are connected [Newman 2004]. Similarly, citation networks involve citations in academic papers, but they are distinguished in that their edges usually point strictly backwards in time, making loops extremely rare [Easley and Kleinberg 2010]. Citation networks have long been used to analyze the impact of individual researchers and journals.

Egocentrical network refers to a subnetwork where a whole network is considered from the perspective of one particular node. Only nodes that are connected to this central node are then considered in the analysis. Such a perspective is often relevant in social networks, and also better supported by available data than analyzing a whole network, as an individual may only have access to data concerning themselves and those they are somehow connected to. The *network level* of analysis, on the other hand, attempts to consider the network as a whole. However, in some cases, defining the bounds of a network and choosing what to include in the analysis of the "whole" network is not a trivial matter. [Provan et al. 2007]

2.3.5 Inter-organizational networks

Inter-organizational networks are much like social networks, but instead of the individual, the focus is on whole organizations. These networks represent patterns of collaboration, dependency and competition that are crucial to the organizations themselves. Understanding one's position in such a network can not only provide an overview of the activity of an otherwise hard-to-grasp whole, but also be a distinct competitive advantage. For businesses, inter-organizational networks represent competition, cooperation, suppliers, knowledge transfer, influence and social capital [Basole 2009, Barringer and Harrison 2000]. In fields such as knowledge management, such networks are also called "ecosystems" to reflect their evolution and adaptation to their environment [Basole 2009].

Barringer and Harrison [2000] note that the number of inter-organizational ties is growing globally, despite the fact that many alliances fail. They discuss the many reasons why organizations form ties with others, and call for a more unified inter-disciplinary research on the advantages and disadvantages of inter-organizational relationships. Indeed, inter-organizational networks are an inter-disciplinary topic, touching business, economy, bibliometrics, network theory, visualization and so on.

In the case of universities, the previously discussed co-authorship networks can be expanded to the university level. In addition, the networks of universities have been studied from the perspective of the alumni and their industry connections [Rubens et al. 2011] as well as patents [Xu 2010]. Neirynck and Borner [2007] also present a system for managing the data of a research unit, such as information about people, publications and expertise, to analyze and improve its processes.

Conventionally, the focus has been on a partial set of actors in the network – often a central organization and its relationships – rather than the whole network but there is an interest in exploring the network level [Basole 2009]. This poses completely new challenges for data collection, analysis and visualization alike.

Part II

Methods and Materials

Chapter 3

Case Study Setup

Understanding the exploratory visualization process is central to the success of visualization, but the process is difficult to observe and evaluate due to the very fact that it is not well-defined. In order to investigate this process in practice, we conducted a case study involving the visualization of the cooperation network of a Finnish university.

The study focused particularly on three aspects of the practical exploratory visualization process: 1) its relation to the process as depicted in literature and discussed in Chapters 4– 5, 2) the kind of exploratory visualization that is suitable to the given context, and 3) the kind of results that can be generated by such an exploratory process. These aspects roughly correspond with the research questions outlined in the beginning of this thesis.

This chapter describes the context, goals, and methods of the case study. Its results are discussed in Chapter 6.

3.1 Context

Aalto University is a relatively new institution established in 2010 in the merger of three well-established Finnish universities – Helsinki University of Technology, Helsinki School of Economics, and the University of Art and Design Helsinki. As such, new structures and workings are still forming within the university while its schools and departments are creating new forms of cooperation. In this context, the university's Institutional Relationships unit aims to understand the university's cooperation network which consists of diverse external partners. This network is critical to understanding the workings of the university as a whole. Visualization is used as an early tool to explore the scattered and heterogeneous data currently available to direct later data collection and analysis.

Data involving the university's cooperation network currently exists in multiple, unintegrated forms. Typically, departments have their own data caches. Third party publication portals such as Scopus [Elsevier 2013] and Web of Science [Reuters 2013] offer large amounts of somewhat consistent data on publications, which are an important, but not comprehensive source of information on the function of a university. Additionally the libraries of the three old universities have their own databases on publications and projects. Finally, some unified data collection have been already conducted in the form of Linked Data. With the exception of the publication portals, all data available is ego-centric and only includes information about external partners through their relationship with the university. In summary, not one comprehensive data set is available, and the exploration involves several incomplete and heterogeneous data sets.

This visual exploration project was part of a larger initiative to manage Aalto University's cooperation network and unify data collection related to research activities. From the perspective of the university, it had several goals. First goal was to establish what kind data was currently available and what data should perhaps be collected in the future, and what problems there are in regard to data quality. The second goal was to develop initial visualizations of the cooperation network that would help understand and manage the partnerships of the university. Visualization was chosen particularly because it could also be used for communicative purposes, such as to communicate the value of research data in order to drive structural change necessary to unify practices within the university, and to communicate the activities and influence of the university to its affiliates.

3.2 Goals of the study

From research perspective, the aim of the case study was to understand the visualization process in practice. This will allow us to compare the results of the case study to the picture painted by our literature review for a more comprehensive understanding of the process. In particular, we wanted to find out what kind of visualizations are useful in early exploration of cooperation networks when a comprehensive data set may not be available, and what kind of hypotheses and insights may be generated in such a context. This will allow us to explore characteristics of inter-organizational network visualization especially.

We examined the process by developing visualizations with concrete goals together with domain experts. The context is early exploration in an organization where data and expertise in visualization on one hand and expertise

in the domain area on the other both do exist, but are scattered within the organization. We chose this focus so we could see what kind of challenges the adoption of visualization and data analysis faces in such organizations.

Inter-organizational networks are a relevant theme as this type of data is crucial to very different kinds of organizations in many fields, and network theory is a young and developing field with untold potential to support further analysis. Although visualizations were created and evaluated during the case study, the goal of the experiment was not to compare fundamental visual elements or perceptual choices, or to create very refined visualizations. Rather, we wanted to know how these visualizations work as a part of the overall process, and what features in them support the exploration of inter-organizational networks.

Some research has already been done into what kind of ideas and hypotheses exploratory visualization can produce [Yi et al. 2008, Saraiya et al. 2005]. In this case study, we also investigated this question in the context of inter-organizational networks by collecting and analyzing hypotheses during the visualization process. As discussed earlier, fully understanding the capacity of a visualization to produce insight requires longitudinal study where users have a chance to get familiar with the visualization, but the general characterizations we obtained are still descriptive of this early stage of exploration.

3.3 Data used

Unlike often is the case in visualization, the data set in this case study was not clearly defined in the beginning. The first visualizations were based on several data sets. From these, eventually one data set was chosen to be the basis of the final visualizations.

The first data sets used were very heterogeneous. They included data on publications from the publication database Scopus [Elsevier 2013], financial data from a particular department, and data about research projects from two different sources. The data sets had various known problems many of which were a result of the changing structure of the university and the names of its various parts. Most of the data was focused on three of the schools of the university that had originally constituted Helsinki University of Technology. In many cases, changes in the names of laboratories and departments caused difficulties even in data from before the merger.

Linked Open Data from Aalto's own SPARQL endpoint (`http://data.aalto.fi`), a separate recent development in data management at the university, took a supportive role. In particular, current data about the structure

of the university, including schools and departments, was available in this format.

The final data set consisted of over 3000 records of research projects from 1996 to the beginning of 2012 from the database of the university library ("TKKTutkii"). The database was originally that of Helsinki University of Technology, one of the universities that formed Aalto University, and mainly contained projects of the three schools that originated in it. The records included fields such as title, keyword, year(s), department, and affiliates participating in the project.

Finally, a data set containing billing records from the years 2010-2012 was used as a comparison point to the research oriented data. This data contained financial information about projects and affiliates from the given time period. Although it used the same project numbers as the research data, conventions regarding naming affiliates and defining projects different to the extent that we did not attempt to combine these data sets.

3.4 Methods and tools

The case study consisted of 3 iterative visualization rounds where an overall task was given, data gathered, and two or more visualizations created. These were then compared in a qualitative interview with domain, experts who were interested in and knowledgeable in the data domain and who were also allowed to discuss the visualizations freely. The discussion was recorded and analyzed, with comments regarding the visualizations themselves, the visualization process, and the themes explored in the visualizations (such as partnerships over time) collected. The goals of each round were agreed upon with the domain experts, and visualizations were developed based on these general goals.

The interviews conducted on each visualization round form the core of the case study. Like Heer and Boyd [2005] and Neumann et al. [2007], we wanted to observe natural interaction with visualizations rather than users performing specific tasks. The method used was semi-structured interviews involving the same three participants on each round. In addition, Saraiya et al. [2005] argue that for measuring insight (rather than the relatively objective perception), having motivated users who know the data is critical. Thus we chose as participants in these interviews three domain experts: two members of the Institutional Relationships unit who were passionate about developing the management of the cooperation network, and the head of the Department of Media Technology with a more research-oriented perspective. Each interview lasted for 30-45 minutes.

Questions asked in the interviews were quite open-ended and encouraged the participants to freely comment on the visualizations themselves as well as the themes around them. This made it abundantly clear how differently the data experts saw the meaning data from someone who was familiar with networks in general but new to the particular topic of university and its affiliates. Often, discussions between interviewees involving a topic relevant to the visualization in question (such as the appropriate way to measure partnership strength) spontaneously occurred. Sometimes we had to ask for clarifications of terms and university practices. For all visualizations, we asked or otherwise made sure we obtained the answer to the following questions:

1. What does this visualization tell you?
2. What would you improve in this visualization?
3. What would you like to know that is not present in this visualization?

These questions resulted in answers both very specific to the visualization shown as well as general themes that the experts would like to explore, and so directed both the exploration process and the development of the visualizations. The advantage of this interview methodology was that we gained insight into the context in which the experts operate and what they see in the data, as well as being able to iteratively improve the visualizations to match their needs. On the downside, the visualization process was already constrained into iterative steps, possibly affecting our analysis of the visualization process as a whole.

As discussed in Chapter 2, there are is an abundance of visualization tools and frameworks available. In producing these visualizations, we opted for maximum flexibility to be able to compare small differences and implement interaction, and implemented the visualizations in the Processing programming environment [Fry and Reas 2004-2013]. This was largely a subjective choice based on the author's familiarity with the tool and the fact that initially we wanted to be able to present the visualizations online, which Processing enables through JavaScript. In addition, small Python scripts were used in data processing, such as filtering and cleaning up relevant items.

Chapter 4

Visualization Process in Literature

The bulk of research on information visualization focuses on the development of techniques and visualization systems. However, these operate in a larger context of an exploratory process starting from the acquisition and selection of data and ending in, in the best case, an evaluation of the results. Understanding this process as a whole is vital to developing the technology to support it as well as to undertaking useful, practical visualization. In this chapter, we present a literature review of the process of exploratory visualization. Ten models or characterizations from different fields are discussed. A synthesis model based on this review is presented in Chapter 5, where the steps of the process are described in more detail.

4.1 Understanding the visualization process

Exploratory visualization is a time-consuming task with inherently vague goals. It is an iterative process that starts with acquiring and selecting data and ends with new insight, hypotheses, and ideas. Unfortunately, what happens in between is often left to the imagination and creativity of the analyst.

Several authors have pointed out the urgent need to understand this process in order to develop better tools to support it. Amar and Stasko [2004] conclude that current information visualization systems are good at faithfully representing data but do not really support decision-making. Jankun-Kelly et al. [2007] and Callahan et al. [2006] are concerned with the encapsulation, reproduction and sharing of visualizations and the maintenance and provenance of data. Integrated in a visualization system, these kind of features

would immensely reduce manual work and simplify collaboration. Callahan et al. [2006] go as far as to write that "the generation and maintenance of visualizations is a major bottleneck in the scientific process".

Yet, we want to tackle this problem in even boarder terms. In their interview study of visualization and data analysis practices in enterprises, Kandel et al. [2012] describe how dependent the visualization process is in practice on the analyst, their personal working style, and organizational context. They record a diversity of tools and approaches, hap-hazard scripts, unnecessary repetition, lack of communication between analysts and lost intermediate data sets and results. In other words, the visualization process in not confined to one tool, at least in how they are currently understood. It encompasses challenges related to organizational context, metadata, evaluation, cooperation, communication, goals, and more.

This perspective to the visualization process would, besides the benefits related to maintenance of visualizations, lead to further automation of the most tedious tasks, make visualization more accessible to individuals and organizations with no existing workflow for analysing data, and develop visualization as a more credible analytical tool that applies to real problems. On data mining Myatt and Johnson [2009] write that the well-defined process of data mining "helps us ensure that the results translate into actionable decisions". The goal of visualization should be no less ambitious.

4.2 Existing process models

Despite the focus in research on presentation, some meticulous models and characterizations have been developed in previous studies to describe the process of visualization. In addition, models of data mining, scientific analysis and exploration are relevant due to similar goals, particularly to exploratory visualization. This section briefly introduces some existing models and frameworks, grouped loosely under three themes: data mining, visualization, and exploration. The next section attempts to compare and summarize them.

Each model described in this section is summarized in a figure (Figures 4.1– 4.9), including those that are fairly simple in structure, to allow for easy comparison. In these figures, black lines between nodes denote progression or hierarchical structure, while gray lines imply iteration or other nonlinear connection.

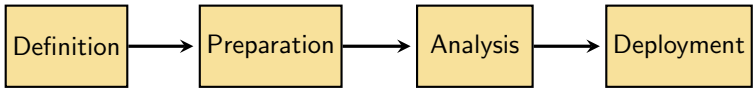

Figure 4.1: The data mining process according to Myatt and Johnson [2009].

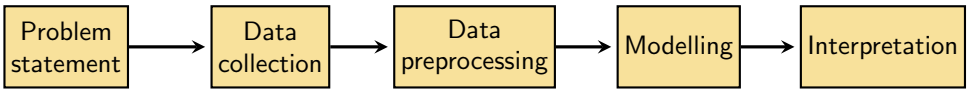

Figure 4.2: The data mining process according to Kantardzic [2011].

4.2.1 Models of data mining

Data mining textbooks, when not structured around the process itself, often describe the data mining process in a few steps such as in Myatt and Johnson [2009]: 1) definition, 2) preparation, 3) analysis, 4) deployment (Figure 4.1). In this simple formulation a a few things are already made explicit that are often not obvious in visualization systems: the need to define goals beforehand, the preparation of data, and the deployment and diffusion of the results to the organization or society at large. A variation is presented in another data mining book by Kantardzic [2011]: 1) state the problem, 2) collect the data, 3) preprocess the data, 4) estimate the model (mine the data), 5) interpret the model and draw conclusions (Figure 4.2). Again, there is a problem statement before anything else and a slot reserved for processing the data, which in visualization is often assumed to be given.

An interview study conducted by Kandel et al. [2012] focused on data analysis and visualization in enterprises. The result is a characterization of the industrial data analysis process and the enumeration typical challenges encountered in it. In addition the authors also describe three distinct data analyst archetypes ("hackers", "scripters" and "application users") that illustrate how differently data analysis can in practice be approached. The analysis process is described as follows (Figure 4.3):

Discovery of the necessary data

Wrangling (or pre-processing) the data into a desired format

Profiling the data to verify quality and suitability

Modelling the data for summarization or prediction

Reporting procedures and insights

Figure 4.3: The data analysis process in a study by Kandel et al. [2012].

Figure 4.4: The knowledge discovery process according to Han et al. [2006].

While the process appears linear, the authors note that in practice it is highly iterative. Overall, data migration between different tools used in different steps was reported a major challenge due to the iterative nature of the process.

Various challenges were reported for each stage of the process. Discovery was considered a significant bottleneck due to difficult access to data, lack of metadata and documentation, out-of-date and conflicting data. Wrangling was also considered tedious, especially in case of semi-structured data such as log files. Questions of data quality were faced in the profiling phase, where some analysts used statistical methods while others relied on intuition. Visualization was used in this context, if at all, in the modeling phase, which also included feature selection (reportedly the hardest part) and typically faced software performance issues when dealing with large data sets. Reporting was considered often incomplete, missing assumptions and information about the original data.

From a more theoretical perspective, Han et al. [2006] describe, in the introduction of their book, data mining as a step in the process of *knowledge discovery*. The process is enumerated as follows:

Data cleaning Removing noise and inconsistent data.

Data integration Combining multiple data sources into one data set.

Data selection Retrieval of relevant data from the database.

Data transformation Aggregating and summarizing the data into forms appropriate for mining.

Data mining Using intelligent methods to extract data patterns.

Pattern evaluation Identify interesting patterns using some measures.

Knowledge presentation Visualization and knowledge presentation techniques used to present the mined knowledge to the user.

An interesting feature of this model is that evaluation is presented as its own step. Mathematical and probabilistic models are used in evaluation. Visualization is mentioned in the final step, however in the book itself visualization is still used throughout the process, for example box plots as early as in the data preprocessing phase to detect outliers in the data. Otherwise the model is very similar to the other data mining characterizations.

4.2.2 Models of visualization

This section introduces three very different ways to characterize the visualization process specifically. Shneiderman's [1996] "information visualization mantra" has become a cornerstone in the design of visualizations and visualization systems. It goes as follows:

> Overview first, zoom and filter, then details on demand.

In the original paper, this mantra is repeated several times, stressing the iterative nature of visualization. While mostly concerned with presentation, we will see that this mantra can also be applied to the whole exploratory visualization process: at first, get an overview of the data, second, zoom and filter to interesting themes, and finally get details about the selected topics.

The *P-Set model of visual exploration*, described by Jankun-Kelly et al. [2007], is one of the few attempts to rigorously define exploratory visualization as a process. It encapsulates visual exploration sessions to enable sharing, storing and analysis of these sessions. An exploration session, in this context, refers to a user session of a visualization software. The model is based on the notion of the *fundamental operation of the visualization exploration process* – the application of a set of parameter values to a visualization transform to generate a visualization result. It presents the major elements of a visualization session as follows (Figure 4.5):

Visualization transforms that take some parameter types as arguments and produce a visual result. Different transforms having the same parameter types and result type (signature) may produce a different visual result.

Visualization parameters with a type (such as colour) and value

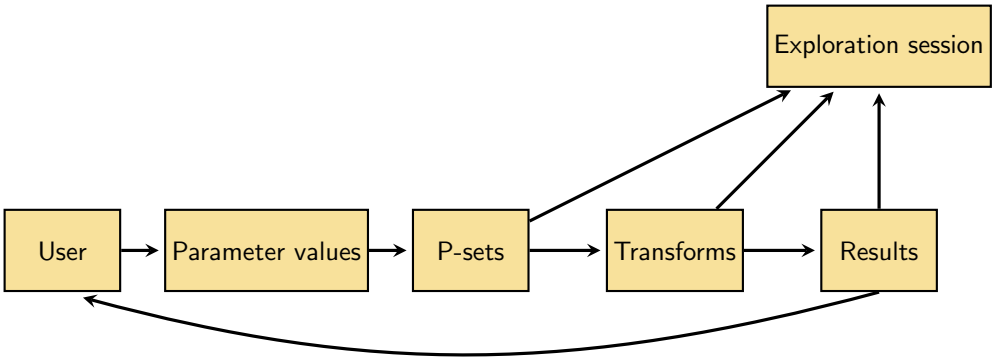

Figure 4.5: The P-set model of visual exploration, adapted from Jankun-Kelly et al. [2007].

P-sets groups of parameter types with values, which are the basis of transforms

Visualization results which are uniquely identified by a particular transform and a p-set

Derivations that contain all the information regarding how a result was created (stored as an "exploration session")

The model is supported by an XML (Extensible Markup Language) based representation format for visualization sessions, and a software framework to manage these representations. While not describing the process directly, these elements show the building blocks of the visualization process and how visual results are created. Describing these entities allows for the storage of visualization sessions, but the authors acknowledge that what the model fails to encompass is the knowledge of the user before and after the visualization session. The perspective is limited to one step of the visualization process.

Neumann et al. [2007], noting the sparsity of literature on the process of visualization, conducted an observation study on the visual information analysis process of groups and individuals. They were particularly interested in how groups engage in information analysis as opposed to individuals. While the study is unique in that it considers both individuals and groups, it did not find major differences between the two. Instead, its main contribution is a framework of eight actions that were observed during the study, where participants were given paper-based visualizations and tasks to solve using them. The actions are as follows (Figure 4.6):

Browse Flipping through visualizations or arranging them in view

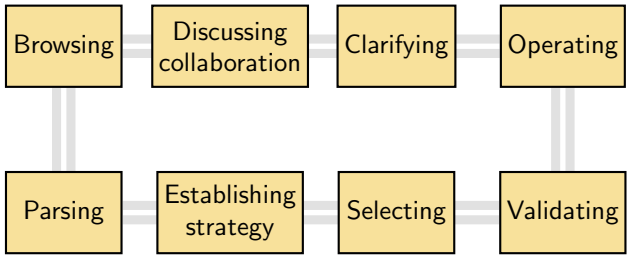

Figure 4.6: The actions in visual analysis by Neumann et al. [2007].

Parse Reading and interpreting a task description, and determining required variables for a task

Discuss collaboration style Only applicable for groups

Establish task-specific strategy Choosing the best way to solve a task using the given data and tools

Clarify data Scrutinising charts and re-reading descriptions

Select data Placing cards in view and discarding others

Operate on data Extracting a value or comparing values

Validate findings Comparing findings to other visualizations, explain to group if available

The study consisted of pre-defined analytical tasks and as such may not be directly applicable to exploration. Nonetheless it is important in that it lists the tasks that form a visualization-based analysis free of the constraints of a particular visualization system. It is notable that when the authors also analyzed the temporal order of these actions, some actions frequently occurred before others (for example, *parse* often preceded *select*), no common temporal pattern emerged. Each group switched between the actions in a different order, with different patterns of repetition and iteration.

4.2.3 Models of exploration

This section introduces three models that involve exploration, navigation and insight, the defining features of exploratory visualization. While not specific to visualization or even data analysis they help us to understand how people approach analysis when the hypothesis is not clearly defined.

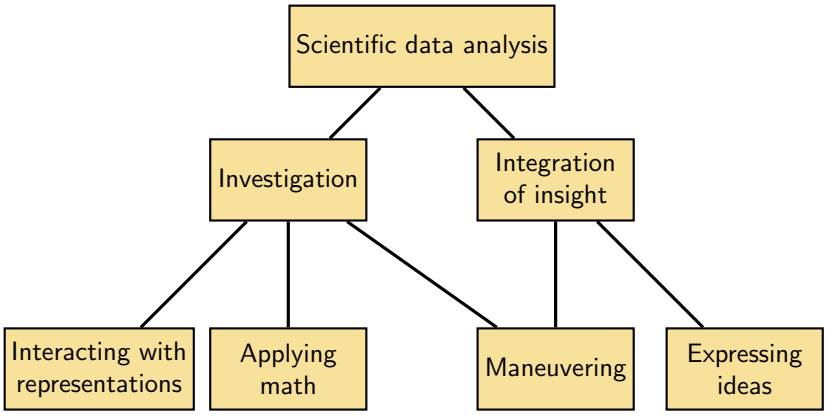

Figure 4.7: Tasks of scientific data analysis by Springmeyer et al. [1992].

Springmeyer et al. [1992] conducted an empirical observation study of scientists from different fields analyzing their own data and presented a characterization of the scientific data analysis process as a result. The model presented lists the tasks involved in the analysis process under two main branches: *investigation* and *integration of insight* (Figure 4.7). These branches contain four sub-branches with tasks related to them as described below:

Interacting with representations includes *Generation of representations*, such as visualizations and tables, *Examination* of visualizations, *Orientation* or basic modification of the visualization such as scaling, rotating and colouring, *Queries* or determining exact values of items in a representation, *Comparison* of elements, and *Classification* of elements.

Applying mathematics includes *Calculations*, *Deriving* new conditions and the *Generation of statistics*.

Maneuvering consists of *Navigation* the visualization that is not concerned with exploration but rather the interface itself, and *Data management* including the transportation and transformation of data.

Expessing ideas includes *Recording* both intermediate and final results and metadata relating to them, and *Describing* results so far, including making judgments about their validity.

The *interacting with representations* sub-branch contains tasks typically associated with visualization, but all of the branches are relevant for exploration. The *applying mathematics* sub-branch might today include machine

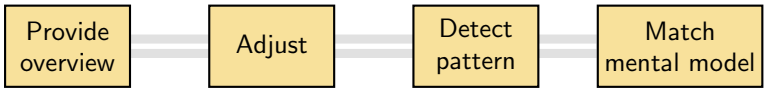

Figure 4.8: The four processes of insight by Yi et al. [2008].

learning and other data mining tasks, while *maneuvering* highlights user interface and data management issues still relevant today. *Expressing ideas* reminds us that communicating final results and recording metadata are a part of the analysis process. The authors were extremely critical of visualization as an analytical tool, and pointed out that many of the tasks are not supported by visualization tools. The original study is a poignant read, and many of its conclusions are still – unfortunately – applicable today.

When discussing exploratory visualization, the concept of *insight* plays a central role. Yi et al. [2008] argue that to understand and evaluate visualization aimed at generating insight, the concept of insight must first be clarified. Based on a literature review, they present four processes through which insight is created (Figure 4.8). The four processes can be described as follows:

Provide overview Understand the big picture of a dataset of interest, find areas to investigate in depth, find out what you already do and do not know.

Adjust Explore a dataset by filtering, zooming, grouping, aggregating, and changing perspective.

Detect pattern Find specific distributions, trends, frequencies, outliers or structures in the data.

Match mental model Reduce the gap in the user's mental model and the data, thereby reducing cognitive load

These processes concretely describe how exploration can achieve insight, and can directly be translated into tasks that a complete information visualization system should support and any exploratory visualization should take into account. For example, *adjust* is a common feature in interactive visualizations, while some explicitly just aim to *provide an overview*. They also hint at the kind of results one can expect from exploration.

In the last model of this chapter, Spence [1999] provides a fundamental framework to describe the concept of *navigation* from the perspective of human-computer interaction. This framework is applicable to physical as well

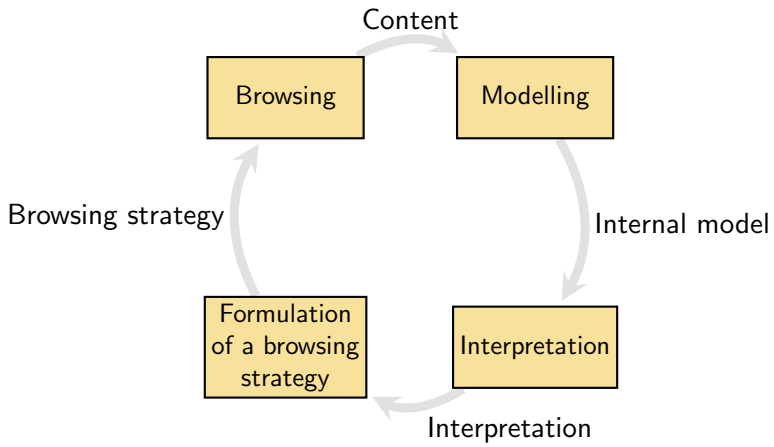

Figure 4.9: Process of navigation according to Spence [1999].

as digital and abstract environments, referring to navigation in the widest sense. Spence uses the word *externalization* to refer to what in our context would be a visualization – a representation of the data of interest, such as a map or a user interface. Navigation is then defined as an iterative process consisting of four steps, each producing a result that serves as a basis for the next step, as follows (Figure 4.9):

Browsing Registration of content in short-term memory based on an externalization. Results in *content*.

Modelling Formation of an internal model, often virtually concurrent with browsing. Affected by existing internal models of the user as well as by the externalization. Results in an *internal model*.

Interpretation Includes deciding whether to continue browsing or not, and what to look for next. Results in an *interpretation*.

Formulation of browsing strategy Particularly influenced by the *affordances* provided by the externalization. Results in a *browsing strategy*.

This model is very abstract, but represents the fundamental steps of navigation, which we here interpret to be roughly equivalent to exploration. It stresses the iterative nature of this process as well as the importance of the existing understanding (mental models) of the analyst and interpreter.

4.3 Summary and comparison

In the previous section, we introduced ten models relating to data analysis, visualization, and exploration. Few of them focus explicitly on exploratory visualization. Yet a comprehensive picture of the exploratory visualization process is not available in literature, despite many visualization techniques and software having been developed. These very different perspectives allow us to make a few observations however. Table 4.1 summarizes all models in the order they were presented.

Table 4.1: Summary of the models and frameworks presented in this chapter.

Author	Category	Perspective	Step order
Myatt and Johnson [2009]	data mining	data mining	consecutive
Kantardzic [2011]	data mining	data mining	consecutive
Kandel et al. [2012]	data mining	enterprise data analysis	iterative
Han et al. [2006]	data mining	knowledge discovery	consecutive
Shneiderman [1996]	visualization	visualization	iterative
Jankun-Kelly et al. [2007]	visualization	visualization software	iterative
Neumann et al. [2007]	visualization	visualization	unordered
Springmeyer et al. [1992]	exploration	scientific data analysis	unordered
Yi et al. [2008]	exploration	insight	unordered
Spence [1999]	exploration	navigation	iterative

The models that describe data mining are each phrased quite differently, but all describe a similar process consisting of a number of consecutive or iterative steps. The overall structure of the process is similar in all of them and can be summarized as:

1. Problem statement
2. Data discovery
3. Data preprocessing
4. Modeling
5. Evaluation
6. Deployment

It is notable, however, that discovery is explicitly mentioned only in the model of Kandel et al. [2012], which is one of the few empirically backed of the models, although Han et al. [2006] also mention data integration. Han et al. [2006] is also the only one mentioning evaluation of findings ("patterns") in their knowledge discovery process.

Models focusing on visualization are more varied. Jankun-Kelly et al. [2007] take the software developer's perspective to visualization, and focuses on the steps a user takes during a "visualization session". For our purposes, these entities describe a specific part of of a more general visualization process in detail. The mantra by [Shneiderman 1996] ("Overview first, zoom and filter, then details on demand."), on the other hand, is focused on user interaction. However, the model by [Yi et al. 2008] which enumerates four processes of insight, uses a surprisingly similar wording ("provide overview - adjust - detect pattern - match mental model"). This perhaps hints to the connection of visualization and insight – that the usual user tasks with an interactive visualization are, in fact, related to different ways of gaining insight.

Neumann et al. [2007] discussed abstract tasks of a user when using a visualization to fulfill a specific task. Their model is directly influenced by the framework of navigation presented by [Spence 1999] which is iterative and is not far removed from the data mining process. However Neumann et al. [2007] point out that the users they observed did not follow any clear temporal structure in their actions. Springmeyer et al. [1992] also characterizes the scientific data analysis process as a collection of available tasks rather than a set of consecutive steps. Like the four processes of insight, this models is not concerned with data collection and processing.

Overall, these models represent very different perspectives but seem to complement each other rather than posing entirely opposing views. Models of visualization and exploration have similarities to those from data mining, but they tend to stress the iterative or even unstructured nature of the process.

Part III

Results

Chapter 5

The Exploratory Visualization Process

In this chapter, we introduce an overall process model of exploratory visualization, based on the existing models from the literature review in Chapter 4. It forms a synthesis of these models from the perspective of exploratory visualization, combining multiple perspectives into a coherent description. First, overall characteristics of the model are discussed. In Section 5.2, each step is examined in detail, and tasks and challenges relating to them is described. Finally, key ideas in the model are discussed. Where applicable, we focus on heterogeneous network data in particular, but the model is mostly independent of the underlying data type or exploratory goals.

5.1 Characteristics

From our literature review it is clear that visualization is not necessarily done by one person alone. In the context of an organization, whether it is a business, a university or something else, several people are directly or indirectly involved in the visualization process. The creation of visualization is usually done by a person who is familiar with visualization tools and techniques, although Neumann et al. [2007] argue that multiple people cooperating on the visualization might be beneficial as data sets keep growing. The *visualizer* may or may not be familiar with the domain of the data in question, and may refer to *domain experts* for ideas and perspectives. Other analysts may also be consulted to for ideas, tools, intermediate results and data leads, although Kandel et al. [2012] reported such cooperation to be fairly uncommon in practice. Kandel et al. [2012] also describe the importance of various *data gatekeepers* such as the IT department, who may be the only ones with

access to or understanding of the data available. Finally, *consumers* of the exploratory results, such as decision-makers or the organization at large, participate in the process.

The process is described here as consecutive steps. However, the temporal order of the process is not clear-cut. It may be iterative and include going back to previous steps [Myatt and Johnson 2009] or the order of the steps may vary entirely across visualization sessions [Neumann et al. 2007]. In the step descriptions, we discuss the likely position of each step in the process.

In this description, we have paid less attention to statistical and mathematical modeling and other more complex tools of analysis. This is partly to differentiate visualization from data mining and partly a result of our focus on early exploration, where such complex modeling may not yet be justified. This is not to say, however, that visualization and data mining cannot exist side-by-side or that modeling would not be a useful component of the process in practice, even in exploratory analysis.

5.2 Steps

In the following, we describe the visualization process step by step. As noted above, the steps are not necessarily consecutive, but they are described here in the most logical order. For each of the steps, we overview its main tasks and challenges. Especially the discussion of challenges in each step draws heavily from the empirical study of Kandel et al. [2012]. In the beginning of each description, we state which models from our literature review are related to the step in question.

5.2.1 Problem definition

Exploration has by definition vague goals, but that does not mean it has no goals at all. The step of stating the problem and defining the goals of the exploratory process is explicitly only mentioned in two of the models in our literature review, both waterfall-like data mining models [Kantardzic 2011, Myatt and Johnson 2009].

In the case of exploration, visualization is only a tool of exploration and perhaps also a tool of communicating findings, but it is not a result in itself. In discussing the data mining process, Myatt and Johnson [2009] state that the problem definition phase should result in a project plan that contains objectives, deliverables, and roles and responsibilities of those involved. These should all be considered in detail sensible to the project in question. The deliverables may be a report or in some cases simply the visualization itself.

The goal for exploration is not necessarily very clearly defined [Saraiya et al. 2005]. It can be a general theme or topic, or alternatively it could also lean on the four processes of insight by Yi et al. [2008]. Is the aim to *provide an overview* of a topic or a data set, or is it to *detect patterns* related to some specific phenomenon?

While it is natural to assume that the problem definition is the starting point of the visual exploration process, it may not always be so. It is possible that a data set is the spark that begins the process and leads to a problem definition. This step can also be iteratively returned to and the goals refined as the exploration proceeds [Neumann et al. 2007].

5.2.2 Data discovery and selection

Data is the substance from which all visualizations are built. A coherent data set is the starting point of visualization, and often it is assumed to be given. This may be true in some contexts, but often one must start by discovering and then selecting a suitable data set. Data mining literature dedicates a good deal of space to the discovery, selection and processing of data [Kandel et al. 2012, Han et al. 2006].

Discovery refers to mapping what data is available, how to access it, and what it contains. Its difficulty depends on the situation – when a data set is given and clearly documented, it may be a trivial matter. More often than not, however, it is a tedious and time-consuming part of the data analysis process [Kandel et al. 2012]. *Data selection* involves choosing a part of the available raw data that is relevant to exploration goals and retrieving it [Han et al. 2006]. For example, it may make sense to leave out data that is very old or focus on a particular perspective, or it may even be necessary to sample the data because it is simply too big for the software or techniques in the following steps to handle.

Data discovery and selection involve many challenges depending on the nature of the data. Data mining often assumes an integrated, well-maintained collection of databases (called a *data warehouse*) [Kantardzic 2011], in which case the main issues are access and retrieval of the data. In reality however, data may come from different, heterogeneous sources such as public data sets, different databases, logs and so forth [Kandel et al. 2012].

Kandel et al. [2012], in their observational study of data analysis in enterprises, note two main challenges in this step. One is finding data, which is hindered by access rights and lacking documentation. They discovered that finding data is a social process where the data itself and understanding of it is scattered in the organization and it may be more useful to know who knows about the data than where it is. The other main challenge is the

quality of data. For example, fields may be out-of date, too cryptic to understand, or conflicting. All data has some errors and incompleteness, so there is also a trade-off between discarding interesting data and risking making a visualizations that lead to the wrong conclusions.

Because of its tediousness, it is logical to assume this step is not repeated unless necessary. Some additional data may be fetched later if the exploration seems to require additional information about some new facet, but current literature does not provide hints as to how often this might happen.

5.2.3 Data preprocessing

Data preprocessing includes steps like cleaning the data and *wrangling* it to the required format [Kandel et al. 2012]. Like data selection and discovery, data preprocessing is mainly dealt within data mining literature [Han et al. 2006, Kantardzic 2011, Myatt and Johnson 2009].

Raw data is rarely useful as such. Frequently quoted as the most tedious part of data analysis or visualization [Kandel et al. 2012], data pre-processing includes removing entries with missing or erroneous values, aggregation, and feature selection. It is often also simply needed because a given tool or software to be used in the following steps requires the data in some particular format [Kandel et al. 2012]. Two other tasks are also included under this heading. *Profiling* means checking the data quality, familiarizing yourself with the data set, making sure your assumptions about it are correct, and discarding any data of dubious quality [Kandel et al. 2012]. *Integrating* data sets is needed if relevant data is available in several sources [Han et al. 2006]. In some cases, preliminary modelling in the form of aggregation, feature selection and filtering also already happens in this stage [Kandel et al. 2012, van der Heijden 2009].

According to the study by Kandel et al. [2012], the main challenges of data wrangling are the ingesting of semi-structured data such as log files and data from 3rd party services. The challenges of profiling are missing, erroneous and extreme values, missing observations, and wrong assumptions by the analyst. These assumptions would be aided of the data was well documented, but this is rarely the case.

Data preprocessing always follows data discovery and selection and is required to happen before analysis or visualization, so its temporal position in the process is quite predictable. If quality data is deemed insufficient, it may be necessary to go back to the data discovery or even problem definition phases.

5.2.4 Creation of visualization

By far the largest and perhaps the most complex step of the process, the creation of visualization replaces or encompasses the *modeling* and *analysis* steps of the data mining processes in our literature review. Creation of visualizations is the heart of the visualization process and the focus of most research. It is explicitly discussed by Jankun-Kelly et al. [2007], Shneiderman [1996] and Neumann et al. [2007].

The creation of a visualization (or several visualizations) includes explicit or implicit modeling of the data, the creation of a mental model of it [Spence 1999] and then finally also creating a visual presentation, the result of the analysis. These tasks can be highly intermingled as the visualizer tries out different views and approaches to the data. Even with a systematic approach, visualization is highly iterative [Shneiderman 1996, Neumann et al. 2007]. Explicit modeling may include statistical modeling or feature selection [Myatt and Johnson 2009]. For networks, modeling may include calculating basic network statistics such as betweenness centrality and average path distances [Easley and Kleinberg 2010].

The actual workflow of this step is greatly influenced by the tools and techniques selected by the analyst. As discussed in Chapter 2, visualization can involve using specialized visualization software that offer certain techniques and templates, or it can mean creating a visualization from scratch. Many tools also exist for specialized fields such as biology, social networks, bibliometrics and so on. In any case, the visualization is produced by certain transformations (techniques) using certain sets of parameters and values as an input, creating a visualization result [Jankun-Kelly et al. 2007]. This resulting visualization is then placed under further examination in the next step.

Challenges of the visualization step are likewise highly dependent on the tools used. Some tools may be limiting in the possibilities they offer, while on the other hand these limitations reduce the confusion resulting from the abundance of tools and techniques available and the lack of clear indicators of when to use which technique. Comparison of several visual presentations of the same data may be necessary before a satisfying result is found [Neumann et al. 2007, Callahan et al. 2006]. We leave visualization-specific challenges such as those related to space, dimensionality and the balance of overview and detail out of the scope of this discussion. According to [Kandel et al. 2012], the hardest part of the modeling part is feature selection. They also point out that some visualization tools have a limited ability to handle big data sets and sometimes sampling may be required due to software limitations.

Visualization is only possible after data has been cleaned and it is fol-

lowed immediately by analysis. Iteration between visualization and analysis is typical of exploratory visualization [Spence 1999, Jankun-Kelly et al. 2007]. However, sometimes in this phase also it may become evident that more or different data is needed. In addition, quick standard visualization can be used already in the data profiling phase to find out the statistical qualities of the data, and to detect inconsistencies [Han et al. 2006, Myatt and Johnson 2009].

5.2.5 Interpretation

The interpretation or user interaction step is quite unique to visualization in that in it, perhaps even more explicitly than in data mining, the construction of a model (presentation) of the data and its interpretation are distinct and may even be performed by different people. The interpretation step is present in the data analysis models of Springmeyer et al. [1992] and Kantardzic [2011] and discussed in depth by Shneiderman [1996] and Yi et al. [2008].

Whether the visualizations produced in the previous step are interactive or not, they will involve interpretation by the viewer. Especially in exploration, the aim is not only to present the data but to produce insight. This may happen as the analyst who produced the visualization explores the result in detail [Neumann et al. 2007, Springmeyer et al. 1992] or by the active input and interpretation of another viewer, for example an expert in the domain of the data. This approach is supported by Neumann et al. [2007] who point out that as data sets keep growing, groups of analysts may be more useful than individuals. In some visualization software, this step is highly intermingled with the step of producing visualizations, while in complex interactive independent visualizations they are most distinct. Heer and Boyd [2005] also describe a situation where exploration of a visualization of a social network results in spontaneous collaboration and story-telling by users and bystanders.

To understand this step better, Yi et al. [2008] propose four distinct processes through which a user gains insight. Their paper is largely focused on users using information visualization systems, and, implicitly, expert data analysts, but the visualization mantra by Shneiderman [1996], worded quite similarly, applies to the general user of a visualization. The message in both is that first an analyst needs an overview, then they zoom into some area of interest, and finally fetch detailed information about some instances of data. Many authors [Spence 1999, Yi et al. 2008] also note that the prior knowledge and understanding of the data of the user affects interpretation.

As the ideas produced by user-interaction are subjective and may be hard to put into words, documenting them is a challenge. User interface related

issues also arise in this step as the visualization may not be as approachable as its constructor thought [Shneiderman 1996, Heer and Boyd 2005], resulting in difficulty in interpreting the visualization or wrong interpretations.

Interpretation follows directly, or is mixed with, the visualization step. Based on the interpretations, the exploratory process can go back to reiterating any of the previous steps, such as finding more data to answer new questions or making new visualizations.

5.2.6 Evaluation

The step of evaluating the results of an exploration was not present in most of the models in our literature review, but was mentioned by Neumann et al. [2007] and Han et al. [2006]. In addition we draw on literature on the evaluation of exploratory visualizations in general.

As Ellis and Dix [2006] point out, visualizations are hard to evaluate because they are not valuable in themselves but for the results they yield. This is particularly true for exploratory visualizations. While the evaluation of visualization techniques is important for the development of visualizations and visualization software, in the general process of exploratory visualization it is more important to evaluate the *findings* – the insight and hypotheses generated in the previous step. Yi et al. [2008] point out that factors such as user motivation, usability and perceptual issues such as clutter affect insight. Spence [1999] discusses the existing mental models that affect any interpretation. Interpretation, in other words, is subjective and only partly dependent on the visualization, and should be evaluated separately.

Literature on evaluating the results of exploratory visualization is scarce. We propose that insight can be evaluated from two perspectives: its *truthfulness* and its *usefulness*. The usefulness depends highly on the organization but some metrics should be set up so that the usefulness of the whole process can be examined. The truthfulness can be evaluated relative to the data using statistical methods [Han et al. 2006] or some other reference point. In the study by [Neumann et al. 2007], informal evaluation was performed in groups by comparing visualizations and by one individual explaining their reasoning to the group and convincing them of their interpretation.

Evaluation can only happen after interpretation and should ideally precede deployment of ideas. It is not likely to be repeated unless the whole process is repeated.

5.2.7 Presentation

The final step in the process has many names in literature. Springmeyer et al. [1992], given a scientific context, call it *expressing ideas*, Kandel et al. [2012] talk about *reporting*, Han et al. [2006] use the term *knowledge presentation* and Myatt and Johnson [2009] use the mechanical *deployment*. In the context of exploratory visualization these all however refer to the same thing – diffusion of the ideas generated to the intended consumers of the insight, be it managers, the organization at large, or the whole scientific community.

If a data was explored just because the analyst wanted to get familiar with it, then deployment is unnecessary. In any other case, consumers of the insights created by exploratory visualization are not the analysts themselves. The most obvious way to deploy the results of an exploratory visualization is the presentation of the visualization itself. Depending on the nature of the visualization, it may be all that is needed. If a single visual end result is not available, or if the visualization is not self-explanatory, presentation includes some form of reporting.

Kandel et al. [2012] note that the most common problems related to the reporting of findings in data analysis are assumptions related to communication, and static reports, which cannot include filtering and the original data. They especially find that assumptions made during the process are rarely communicated efficiently.

5.3 Summary

From the model presented here, we can clearly see that visualization is not limited to the creation of visualizations. Even in exploration, problem definition should not be forgotten entirely. Data discovery and processing steps are time-consuming and unavoidable parts of the process. Exploration and interpretation take place in many steps and levels throughout the process – in fact, interpretation is likely to continue even after the presentation step as the results are diffused within the organization. We also see that the visualization process involves many stakeholders.

In fact, the process is very similar to that of data mining, at least in terms of the steps it consists of. However, we think the key difference is that the visualization does involve an approachable, visual presentation. This makes it easier to share intermediate results and to collaborate [Neumann et al. 2007]. The nature of those visual presentations makes the relationship of the visualization and the interpretation steps quite complex.

On one hand, interactive visualizations make it possible for the visual-

ization and interpretation steps to be clearly separate and even performed by different people. Visualization systems can in a sense perform half of the analysis by imposing certain restrictions on the presentation, to be refined by a user, and finally perhaps interpreted by someone else. On the other hand, visual presentations may only be intermediate results, and interpretation may be very closely intermingled with the creation of visualizations, which are tweaked as the visualizer gains more understanding of the data they are working with.

If all of the steps of the process are mixed and overlapping, the question arises whether it makes sense to think of the process as separate steps at all. However, recognizing the tasks as separate allows us to consider their requirements and analyse their interaction.

Chapter 6

Case Study Results

In the previous two chapters a theoretical model of the exploratory visualization process has been established. In this chapter, the results of the case study whose setup and methodology were outlined in Chapter 3 are presented. First, visualizations produced during the iterative development are presented along with a summary of interview results for each separate cycle, and then overall findings are presented. Discussion of the results of the study in light of the process model in the previous chapter is deferred to Chapter 7.

Quotations in this section have been translated from Finnish, the original language of our interviews. Some of the labels in the screenshots of visualizations have been obscured for reasons of confidentiality so that the name of an affiliate, for example, is replaced with the word "Affiliate".

6.1 Visualization rounds

The case study took the form of an iterative visualization process consisting of 3 rounds. On each round, new visualizations were constructed and presented to our three domain experts – two members of the Institutional Relationships unit and the head of the Department of Media Technology of Aalto University. Each round builds upon the previous one. This section describes each of the visualization rounds separately, presenting the visualizations created and a summary of interview results, revealing the overall progress of the study.

6.1.1 Round 1: Initial example subnetworks

The first visualization task was to take a closer look at two particular subnetworks, one based around an external affiliate (a company with known history

of cooperation with at least one of Aalto's departments) and one around a department of the university. Data wrangling was kept to the minimum at this stage in order to simply get a first overview.

Two distinct interactive visualizations were created, one focusing on the hierarchical structure of the subnetwork, and one on the time dimension of cooperation. Four datasets were used:

Dataset 1 A dump of the research database of the Helsinki University of Technology ("Tkktutkii"), filtered for the department in question.

Dataset 2 Funding-based project listing from the department controller.

Dataset 3 Published papers from the public citation database Scopus [Elsevier 2013] where the company in question was listed as an affiliate.

Dataset 4 Internal spreadsheet data on projects with the company.

An additional data set published as Linked Open Data (available from `http://data.aalto.fi/`) was used to create the university's inner hierarchy for the graph. One more data set was available but was left unused because it lacked documentation that would have helped us to interpret it, a notable example of the importance of metadata. All data sets were small and contained inconsistencies, especially related to affiliates' names (a particular affiliate might be referred to with several names) and department assignment (the names of various departments and the overall structure of the university has changed over the years, creating difficulties of attribution).

Figure 6.1 shows the graph visualization with Dataset 4. Each circle represents a unit of the university, and each number the number of items (in this case projects with the external affiliate) involving that unit in the dataset. The units are presented in a hierarchical order, with the whole university in the middle, schools on the inner circle (yellow) and departments within schools on the outer circle (green). Some departments were not placed under schools and are connected directly to the university. The visualization is interactive - solid circles are *closed*, while circles with a white background are *open*, showing their children. In the screenshot, a school with 9 recorded projects has been opened to find that all of those projects have been with one department.

This is an intuitive representation of an ego-centric network and a straightforward first attempt to understand it. In essence it is a custom hierarchical layout for the traditional node-link diagram, putting an emphasis on the structure of the university and contributions of each of its individual departments. The interactivity allows the viewer to freely examine cooperation on the school or department level.

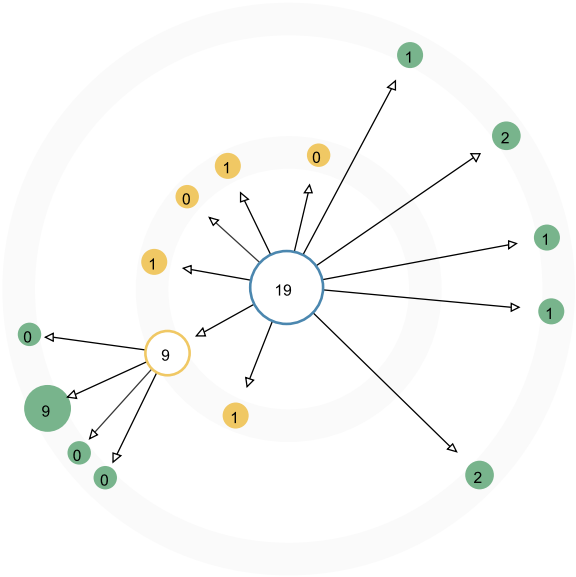

Figure 6.1: Layered graph visualization of Round 1 (labels hidden).

Figure 6.2 shows the time series visualization with Dataset 1. In this visualization, lines represent affiliates, the x axis shows years, and the y axis items (in this case research projects) with that particular affiliate active on a given year. The user can hover over a particular line or affiliate to highlight that one – in the picture, one affiliate has been highlighted. Affiliates with few projects are combined into a category called "Other", and the highest line ("All") represents the sum of all projects, giving an overview of the development.

This visualization was an attempt to present the network from the perspective of evolution over time. In essence, it is a simple time plot, where individual projects have been hidden and instead project count per affiliate is shown on the y axis.

We were positively surprised by the amount of discussion these simple visualizations created in the interview with domain experts. Roughly four types of ideas and insights arose during the interview: 1) hypotheses and themes to be explored in the future, 2) concrete feedback about the specific visualizations, 3) general discussion around the topic of cooperation, affiliation and a university's ties, and 4) realizations about the current state of relevant data and ideas about how it could be improved. The visualizations themselves received both overall praise as well as critique of missing labels and ambiguous elements. Our interviewees were particularly enthu-

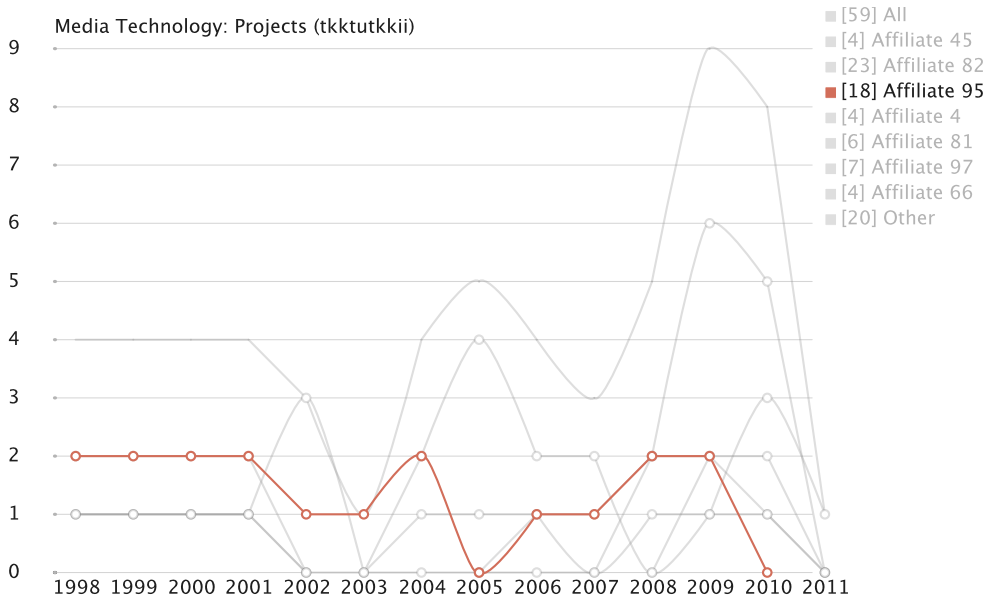

Figure 6.2: Time plot visualization of Round 1, with one affiliate highlighted.

siastic about being able to grasp an overview of the data available in the organization:

> *With this we can tell about the current situation, that the data quality is bad, and then we can enter the discussion about it.*

Exploration-related hypotheses and questions generated by the interview included things like the definition of the strength of a relationship in this context (instead of project or publication counts, for example money, scientific contribution, people and work hours), and the distinction between internal and external partners. Additionally, many potential data fields were identified, some of which were more realistic in terms of actual implementation in the future than others. The need for various metadata such as where the data originates and how it has been collected surfaced several times.

6.1.2 Round 2: Focus on collaboration

Based on the first round, goals for the second round were set together with the domain experts. First, instead of the department level bigger entities such as schools were decided to be more interesting. Second, the whole network should be brought into view instead of an egocentric perspective as much

as was possible with the data that was mostly concerned with the university itself. In practice this meant abandoning explicit focus on particular affiliates. Finally, project or publication count was deemed not a very useful measure of cooperation:

> *From the perspective of the department, rather than the number of projects what is interesting is their volume.*

On this round, only one data set was used – the whole unfiltered research database dump which was in the previous round filtered for a particular department (Dataset 1). It contained over 3000 entries from years 1994–2012 and its fields were quite well documented. This data was used as a basis for three visualizations, each focusing on a theme that had arisen in the first interview. New attributes from the data were used, such as keywords listed for each project, and man months used in each project. Man months became the unit of affiliation measure, instead of project or publication count.

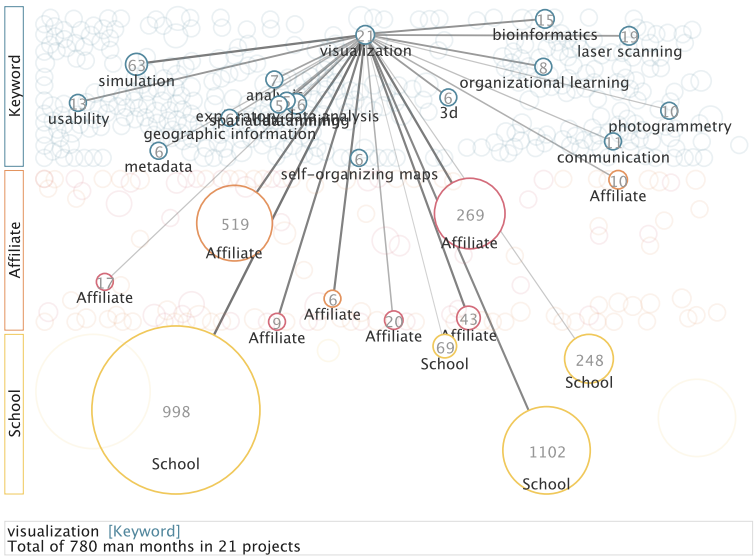

Figure 6.3: Node-link visualization of Round 2, with the keyword *visualization* highlighted.

The first visualization is depicted in Figure 6.3. It is a development of the graph visualization from Round 1, with the structure of the university, familiar to all involved, hidden and instead nodes of the same type confined to the same vertical space (indicated with the labels with node types like

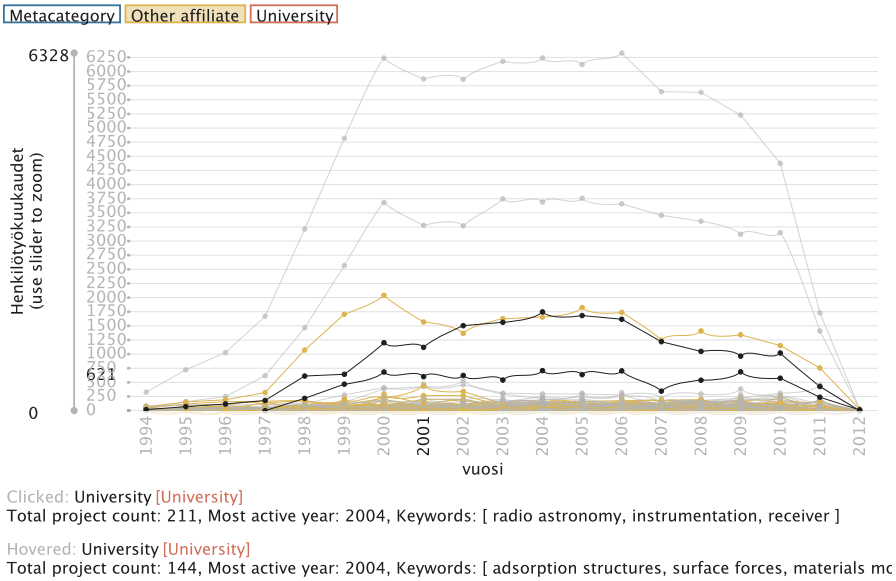

Figure 6.4: Time plot visualization of Round 2, with two affiliates highlighted.

"Keyword" on the left). New types of nodes were brought to the fore. Instead of departments, there are now only schools (in yellow). Affiliates are red, and keywords found in projects are blue. Edges indicate that two nodes appear in the same project. Node radius reflects total man months used in all projects and number in the node denotes the number of projects for comparison.

In the screenshot, the keyword *visualization* is selected, highlighting all the nodes that it is connected to and dimming others. We can see that there have been 21 projects with that keyword, all of the schools but one have done research involving visualization, and that there are several affiliates that are also involved in this research, some considerably bigger (in terms of the size and count of projects) than others.

Since the colour indicates the type of the node, a more traditional node-link layout algorithm could have also been used. However we found that the difference in the nodes' types here is important enough to separate them in the space. In addition, a small box containing information about the selected node was added to the bottom of the visualization, as we noticed that our domain experts were interested in individual nodes rather than the overall structure.

Figure 6.4 shows the second visualization of this round, which is based on time series from Round 1. As before, years are on the x axis, but now the

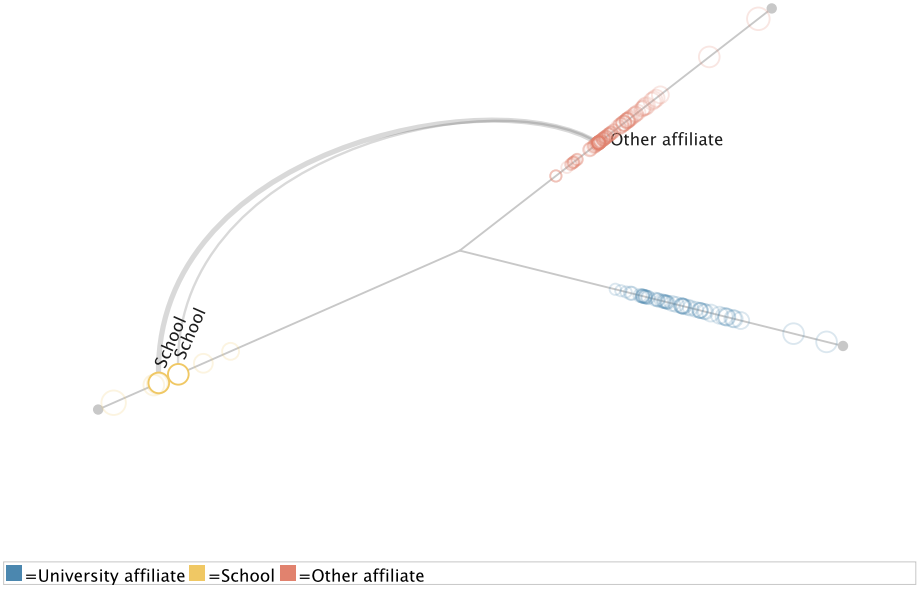

Figure 6.5: Hive plot visualization of Round 2, with one affiliate and its connections highlighted.

y axis represents man months rather than project counts. Affiliates can be highlighted by *type*: university affiliate or other affiliate. In Figure 6.4, one affiliate is highlighted, and the info box in the bottom of the picture provides additional information about it such as the total number of projects and common keywords. Since this view is dominated by the few most prominent affiliates, it is possible to zoom in on the y axis by dragging handles to the left of the y axis. Affiliates can be highlighted by either clicking or hovering, allowing comparisons. Highlighted affiliates are shown in black lines, while coloured lines show affiliate type highlights.

The third visualization of Round 2 was completely new, based on the idea of a hive plot [Krzywinski et al. 2012]. It is shown in Figure 6.5, where one affiliate is highlighted (without highlights, all of the edges are shown, giving an overall pattern of connections). The affiliates here are divided into two types as in the time series, indicated by both colour and the axes of the hive plot. The position of the nodes on their axis reflects the total number of man months of the projects where they have participated, in a logarithmic scale, so that nodes with the most man months are furthest away from the middle.

The idea of the hive plot is to give meaning to the position of the nodes, and this visualization was developed as an alternative to the node-link diagram. From it, an overall pattern of affiliation is easier to see, but nodes

overlap each other and finding a particular one might be difficult.

On this round, detail and labels such as titles and scale indicators were added to the visualizations, and this clearly improved their ability to provoke discussion. The overall pattern of ideas was similar to the previous interview, with the four topics and each visualization all receiving a roughly equal share of comments. The only exception to this was the graph (first visualization) that seemed to create more discussion overall. The interviewees were particularly happy to see keywords, but also wished to see the people in the projects:

> *At the end of the day personal relationships are essential to the birth of new projects.*

The focus of the discussion however shifted from what is a good measure of partnership and general lines of cooperation in the university, to what these visualizations could in fact be useful for. A definite management perspective emerged, with discussion detailing the way partnerships are formed and what kind of partnerships are needed. More ideas about what kind of data would be useful emerged also. Development ideas for the visualizations were markedly more specific, such as the request to include a search box for affiliates.

6.1.3 Round 3: Refinement and additional data

Based on the interviews on Round 2, two main goals in addition to minor refinements for the visualizations were set: the inclusion of data about key people in research projects (which was known to exist in the data base) and the inclusion of financial information as a comparison point to man months as a measure of project size.

A new data set including billing information from years 2010–2012 from all of the schools of the university was provided in addition to the research database from round 2. Thus we unexpectedly had to do some more data wrangling and modeling at this stage of the process to be able to present information that was project specific (man months and keywords) and information that is both project and affiliate specific (billing) in the same visualization.

The time series and node-link diagrams from last round were developed further, while the hive plot was discarded as based on the interview, it proved the least interesting of the three. A search box allowing the search of a particular affiliate, keyword or other node was included in both visualizations at the request of the domain experts.

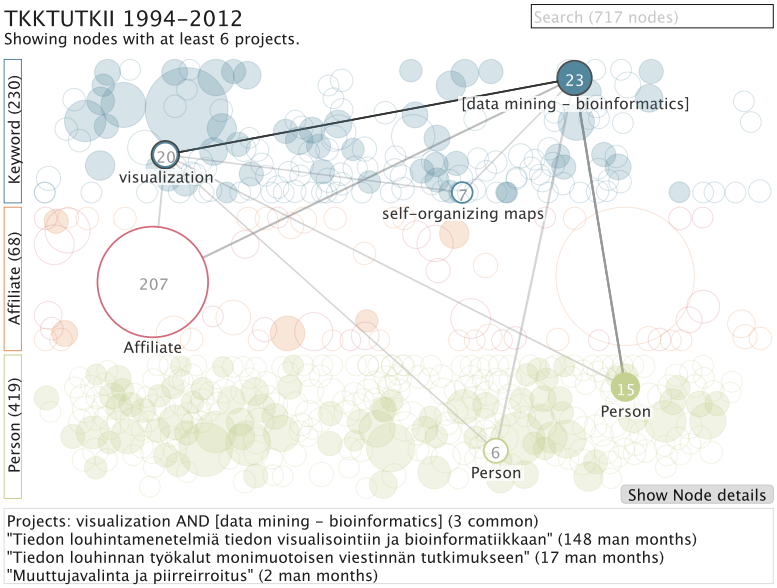

Figure 6.6: Node-link visualization of Round 3, with two keywords highlighted.

The improved node-link diagram, depicted in Figure 6.6, included two major improvements from the previous round besides the search box. Both were based on specific requests. First, our interviewees wanted to include a new type of node, Person, in the graph to see contacts relevant to cooperation. Second, the users wanted a way to select two nodes in the graph and see what nodes they were both connected to (what keywords, affiliates et cetera they shared). In the screenshot, two keywords are selected, showing one person and one cluster of people, a third keyword, and one affiliate associated with them both. Clusters were introduced to reduce clutter and overlap. Clusters are shown with solid circles while simple nodes are filled white. Also, additional data such as listing projects was now shown in the box at the bottom of the view.

We also wanted to include the time dimension in this visualization, allowing the user to specify a time span to examine. However, we did not have time to implement this. It is possible to choose the time span however by filtering the input data, as has been done in the screenshot (only showing data from 2000–2011).

Figures 6.7 and 6.8 show the time series, which underwent similar improvements. In particular, we made it possible to specify the unit on the y axis interactively. Possible units were project count, man months (if available

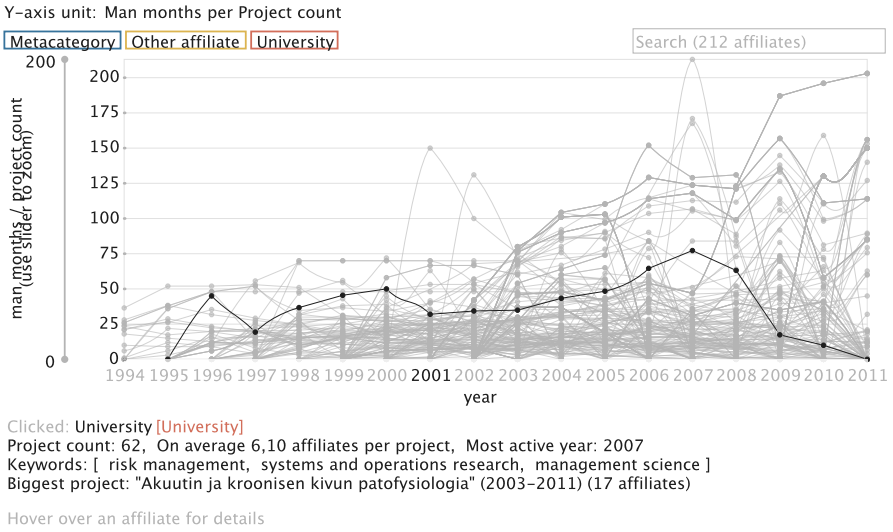

Figure 6.7: Time plot of Round 3 with one affiliate highlighted (project data).

in data), billing (if available in data), and any combination of those, such as man months per project. More detailed data about each individual affiliate was also included in the box below the time series. To handle the large sums that came with the new data sets, a logarithmic scale was introduced. It was automatically used whenever the maximum value on the y axis was large enough, which proved sometimes to be confusing for users.

The new data set, despite only spanning a few years, was fitted to the time plot for comparison with the research data set (Figure 6.8). While the image on first glance looks quite uninteresting due to the data only compassing a few years, the comparison proved useful in that our data experts were very familiar with the numbers involved and could use that to verify their understanding of the visualization. It also sparked discussion around the financial theme in general.

The discussion in the last interview was somewhat different from before. Data was discussed less, and mostly previously stated points were reiterated. There were fewer exploratory themes and more specific questions like *why* this line is falling or *what* are the links between these two nodes:

> *Could it be that the project count has fallen, but there has been bigger, more expensive projects?*

Interest was strongly in specific affiliates and nodes and exact numbers

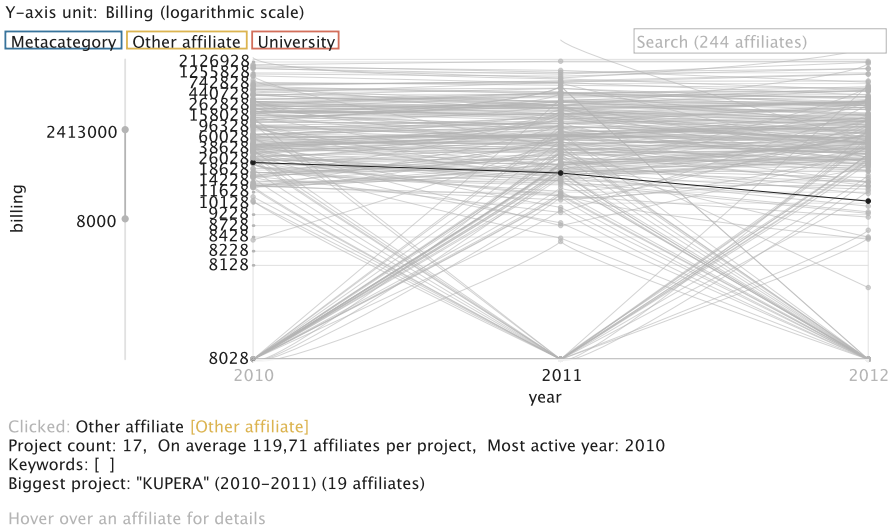

Figure 6.8: Time plot of Round 3 with financial data, zoomed on the y axis.

instead of the big picture. We took this to mean that our experts found the visualizations useful tools in practical exploration that is relevant to their work, and that the previous rounds had covered more general exploratory themes. Another indicator of this was that many of the questions posed were also answered using the visualizations themselves, rather than left for the next iteration to solve. Some themes for further research were still also uncovered.

6.2 Findings

In the previous section, we followed the case study from beginning to end, and showed how very simple visualizations developed into complex interactive exploratory tools. At the same time, more general exploratory topics such as the features of the data and areas of interest for further research were uncovered, and the process of exploratory visualization itself was seen in practice. This section summarizes findings of the whole study regarding the original three aspects of interest in the study: the overall exploratory visualization process, the visualization of inter-organizational networks, and hypotheses and ideas generated by visual exploration.

6.2.1 Process characterization

In this study, the process was constrained to be iterative from the beginning. Despite this, other features of the process can be examined as potentially descriptive of the exploratory visualization process in general. The fact that the process was constrained in time to be about five months long is only realistic in an organizational setting.

The non-linear structure of the process was was clear in the case study. The visualization rounds were quite different, containing different amounts of data wrangling, fixing specific problems, and experimenting with visual presentations. Data collection and wrangling were most present on Rounds 1 and 3. Fixing the user interface problems in the visualizations continued throughout. Exploration and interpretation were present on all rounds and many levels of the process: in the development of the visualizations, the discussions in the interviews, as well as in the interaction of the domain experts with the visualizations.

Data heterogeneity was also a defining feature of the process. Although data was known to exist, it was not clear which data set to use. Very different data sets in terms of source, size, completeness and content were used, sometimes side by side. Data hunting was a social endeavour, involving many forwarded emails and new contacts, and also many clarifications of the exact nature and origin of a particular data set.

Besides the heterogeneity of data, the importance of data confidentiality and other metadata was highlighted. In the case of inter-organizational networks, confidentiality was related to financial data as well as information about external affiliates. This was one of the strings attached to the data that data gatekeepers would know about but that was not part of the data itself. Other metadata that would have been useful included the naming practices involving entities like projects and affiliates.

Finally, an interesting feature of the process was the roles of the different stakeholders. What was imposed in the setting of the case study was that there would be a visualizer, who was not very familiar with the domain of the data, and domain experts, who on the other hand were not familiar with visualization or data analysis tools. Other stakeholders, however, also surfaced, especially consumers of the results, both external and internal to the university. Screenshots of the visualizations were used to communicate value to external affiliates and internal decision-makers alike.

It became abundantly clear that understanding of the data domain was crucial to the process. The visualizations were developed heavily based on the interviews, as the domain experts could see things in the visualizations that we simply could not, and also asked for distinctions to be made that we

would not have thought of, such as the distinction to university and other partners that was made on Round 2. The perspectives of our different experts also complimented one another, when sometimes one of the interviewees could explain a phenomenon that another found in the visualization. This suggests that collaboration is highly beneficial for exploration.

6.2.2 Visualization of inter-organizational networks

The visualizations developed in the case study became highly tailored to the specific context they are used in. Thus, we will not present a comprehensive evaluation of those specific visualizations here – that is best left to when the final visualizations are in their intended use. Instead, we suggest some generalizations based on the things we learned from feedback they received in the interviews.

First, let us make some important observations not specific to the visualization of inter-organizational networks. Usability matters, even in rough, exploratory visualizations. In our study it was particularly important because of the collaboration with the domain experts, who needed to understand the visualizations effortlessly. Over and over again we were asked to clarify elements of the visualization that we thought were obvious, or not important. Clear presentation was requested also because the visualizations could then also be used for communicative purposes. Second, the amount of enthusiasm even for simple visualizations was unexpected:

> *What we do today in practice – we call research support and it takes three days and then we get this incomplete spreadsheet. Compared to that this is from a different planet.*

This is evidence for the need for the captivating and communicable approach to data analysis that visualization offers. Finally, the high level of interactivity gives any visualization a different exploratory dimension, allowing domain experts to find the things that they find interesting.

The single most important analytical theme related to inter-organizational networks specifically in our interviews was the definition and measure of a relationship. Understanding the implications of the project count as a measure, for example, required familiarity with how projects are recorded. We also used man months, since they can be thought to be approximate to the cost of the project, as well as actual sums. There are many more results of cooperation in the university world, however:

> *What can be challenging in terms of data, is that theses and papers are produced in the largest quantities, and there is always*

data about money, but things like patents and new businesses are rarer. In a data set there may only be one or zero such instances.

We also showed that graphs and matrices are not the only way to approach networks. Our conceptually simple time plot was as successful as our node-link diagram in provoking questions and providing relevant answers. Time remained an interesting theme throughout the case study, and the node-link diagram could also have benefited from a time perspective.

Finally, we discovered that at least in the domain of this case study, our users were extremely interested in particular affiliates and relationships rather than looking at an overall structure. This might be due to the fact that the data available only enabled an ego-centric presentation. However, it is safe to say that Shneiderman [1996]'s visualization mantra applies for networks as well and that showing details about nodes and edges is important.

6.2.3 Hypotheses and insight

The goal of exploratory visualization is to create hypotheses and insight. What kind of hypotheses and insight? In this section, we try to answer this question in terms of inter-organizational networks by looking at what kind of things surfaced during the case study. While recording insight is a notoriously difficult [Yi et al. 2008, Saraiya et al. 2005], we base our findings here simply on the thoughts vocalized by our interviewees, either when directly asked a question or when speaking aloud while interacting with the visualizations.

The ideas regarding the cooperation network that surfaced in the interviews can be roughly divided into three types: 1) interesting *themes* that should be explored more, either in visualization or by some other means, 2) specific *questions* arising directly from the visualization at hand, 3) singular interesting *facts* learned from the visualizations. Of these, by far the largest group was the first (21 recorded instances in 3 interviews). It was also the predominant type in the first round, while specific questions and facts began appearing more often towards the end of the study.

We also categorized these ideas according to whether they were covered by visualizations, expertise of the interviewees or left open within the scope of this project. By "covered" we mean for example that a theme was explored in a visualization or explained by a domain expert, or a question was answered directly by a visualization or a domain expert. *Facts* (points) are all "covered by" visualization since they arose directly from the visualizations themselves.

Insights from the case study are summarized according to this typology in Figure 6.9, where the size of each rectangle reflects the number of insights of that type (total count is 46). For example, in most questions and themes

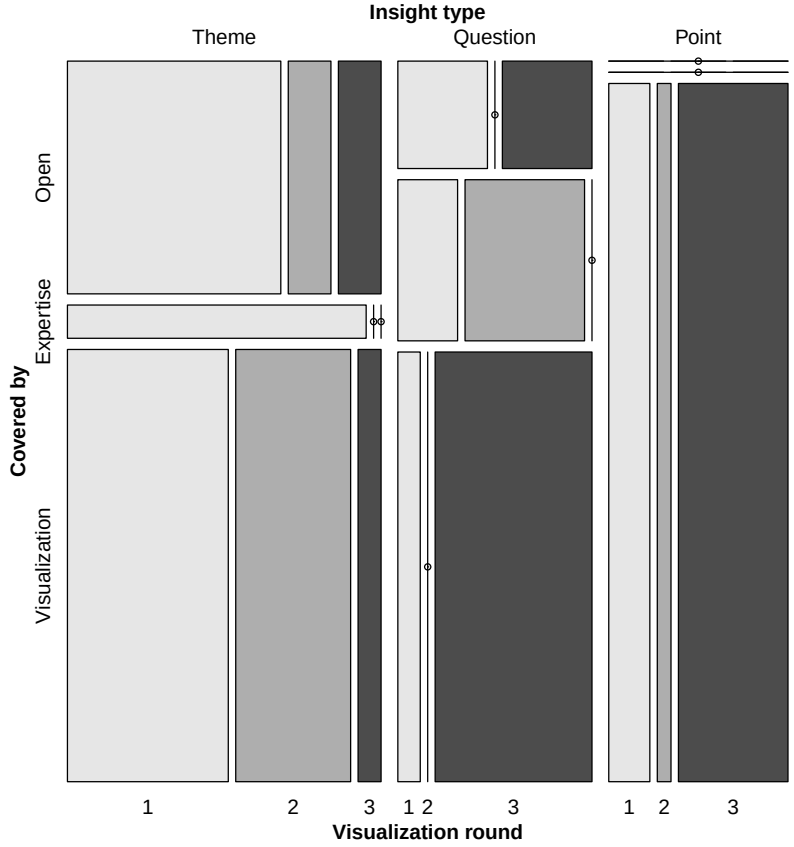

Figure 6.9: Mosaic plot summarizing insight types from the case study.

were covered by visualization, while all of the questions asked on the second round were covered by expertise. In the following, we give examples of each category.

Themes did not go deep into network theoretical questions but were rather specific to the network and context in question. Although in the last section we stressed the importance of showing detail in visualizations, the more general themes were considered by our interviewees a valuable result of exploration. Many of these themes were explored in the next round of visualizations, such as keywords and people, but some were left unanswered during this study. For example:

> *Here too the question arises that who has been involved and if there has been one important contact at the company or one person or unit that works with them.*

Specific questions began to be asked as the visualizations grew more complex. They involved exact numbers or trends, such as whether cooperation with a particular affiliate has slowed down in recent years. These questions were generally answered by visualizations directly, although some were left open. This indicates need for further development, as these were the kind of questions that were evoked by the visualizations directly and should have, ideally, been answered by them as well.

Facts were simple observations, such as the effect of the global economic trend on projects or just pointing out the biggest affiliate in a certain context. They were often direct answers to questions, but sometimes also spontaneous observations. For example:

> *It seems that the cooperation with [a particular affiliate] is a fairly recent thing.*

What is not included in the summary above are insights to the current state of data in the organization. Since one of the exploratory goals in our case study was to get an idea of the data available in the university, these were considered particularly important by our interviewees. However, these kind of information will almost surely surface in any data-based exploration, and is useful for the development of data analysis within an organization.

The insight created directly by our visualizations, and the facts and the questions that were answered by them, were mostly related to particular organizations, their attributes, and their relationship to the university. This is partially because these are the things that interested our domain experts, but also likely a result of the ego-centric perspective of the study. Such details are what are interesting about inter-organizational networks in a certain context, but also what is most readily available in data.

In summary, while the exact categorization of insights in any particular case may be subjective, the general trend seems to be that exploratory visualization yields general thematic ideas and insights about specific patterns. Visualization is capable of providing answers to both of these, but some are likely to always be left open in any project with a realistic scope. Such open questions and themes of interest are still valuable finds for exploration as they lay foundations for future research and analysis. In terms of Yi et al. [2008]'s four processes of insight, exploration produces insight by *providing an overview* of an inter-organizational network and *matching the mental model* of domain experts, while some specific *patterns* may also be detected.

Part IV

Synthesis

Chapter 7

Discussion

So far in this thesis we have examined the exploratory visualization process from two angles – through a literature review and a case study. In this chapter we will combine these results and return to our research questions:

1. What steps does the exploratory visualization process include and what are they like?
2. How can this process support the analysis of inter-organizational networks?
   (a) What kind of visualizations are useful in this context?
   (b) What kind of ideas and hypotheses can exploratory visualization produce about such networks?

The first question is discussed in Section 7.1 where we examine the validity of the process model derived from literature in light of the results of our case study, and also discuss design implications for information visualization systems and practical information visualization. The second question is covered by Section 7.2 where we consider the exploration of inter-organizational networks in terms of visualization design and exploratory results.

7.1 Exploratory visualization process

To understand the exploratory visualization process and answer the first research question we conducted a literature review and synthesized a process model from the literature. We then conducted a case study involving the visualization of the cooperation network of a Finnish university. Here we combine these results and discuss overall implications to anyone involved in visualization, whether in designing visualizations or using them.

7.1.1 The process model in practice

The setting of our case study was early exploration within an organization that has access to data, but whose databases are not entirely coherent, or analysis practices well-established. The progression of the study was partially determined by the context and factors such as our choice of tools and the skills and knowledge of those involved. On the other hand, the literature-based process model we introduced in this thesis is very abstract and general. Here, we take one practical example, our case study, and see how our model applies to its visualization process.

Overall, while an iterative structure was imposed on the process in the case study, the highly non-linear nature of the process, as suggested in our model, was very prominent. New data appeared even on the last round, prompting new iterations of data selection and processing. All steps of the model were present in our case study with the exception of evaluation. While visualizations themselves were evaluated in the interviews, there was no formal evaluation of exploratory results. The separation of the visualizer and the domain experts was imposed by the study setup but other stakeholders as described in our model also became involved in the process.

In Chapter 5 we noted that the process described by our model is very similar to that of data mining. Based on the theoretical model, we suggested that visualization supports the presentation of ideas better than data mining, offers different possibilities for collaboration, and changes the nature of the interpretation step. In the case study, however, we also found that the results are very different from those of data mining. Instead of statistical models or precise theories, they were themes of interest and questions and facts about specific features of the network and its evolution.

Concerning specific steps of the process, the following things can be highlighted from the case study:

Data discovery and selection was a very social process, and heterogeneous and even incomplete data yielded insight.

Data preprocessing , while quoted in literature to be extremely tedious [Kandel et al. 2012], in the exploratory context of our case study was given less attention. Entirely cleaning up the data of duplicate partners for example was considered to be "not the point of this project".

Interpretation occurred on several levels, and was very collaborative. User interface related concerns were the biggest hindrance. Use of incomplete data sets was encouraged and domain experts, in this context,

were able to take problems in the data into account when interpreting visualizations.

Evaluation was, as stated, limited to the interview sessions.

Presentation happened after the interviews by presentations of both the visualizations themselves as well as results in the form of reports.

7.1.2 Design implications

While the steps of the process are intermingled and overlapping, they seem to characterize what really goes on in exploratory visualization well enough to allow us to recognize important tasks and challenges related to it. There is much more to visualization that just making visualizations, and taking this into account would benefit both designers of visualizations and those wishing to use visualization as an exploratory tool.

For designers of (exploratory) visualizations and information visualization systems, recognizing the full visualization process leads to the following design guidelines, which may or may not be applicable to a particular software depending on its scope:

1. Support less than perfect data. Data discovery and processing are bottlenecks of the visualization process, and gaining an overview of various data is especially important in exploration.

2. Show deficiencies in data. Domain experts can take data issues into account, but only if they are transparent enough.

3. Support migration of data between tools. The process is rarely linear and is likely to involve several software.

4. Support metadata, such as information on confidentiality and ownership of the data. This will allow users to gain better control of the process.

5. Support the visualization and interpretation steps as separate tasks. Provide details on demand. Enable collaborative interpretation, at simplest by allowing shareable immediate results such as screenshots.

6. Support evaluation of insight and presentation. Enable the exporting of the exploratory results in some format or another. Even better, help keep tabs of insight during exploration and present a summary of them.

For those who plan to engage in exploratory visualization, such as analysts or managers, it is important to notice the social aspects of data discovery and plan accordingly. While not covered in great detail in this thesis, the huge variety of visualization tools enables very different approaches to visualization depending on the goals of the visualization, skills of the analyst, and the data available. For example, while custom visualizations such as those produced in our case study may be able to address very specific themes, simple tools combined with a well-planned interpretation step could provide interesting results much quicker. One can also benefit from considering all the different levels on which interpretation can happen (user interaction, iteration, and discussion). Finally, evaluation and presentation steps should not be forgotten, even though it depends on the context how they can be implemented.

7.2 Visualization of inter-organizational networks

In this study, we have focused especially on the exploratory visualization of inter-organizational networks. It would seem that the data type (or types) involved is not a significant factor in the overall structure of the visualization process. However, contents of specific steps may be very different depending on the topic of exploration. To answer the second research question ("How can the visualization process support the analysis of inter-organizational networks"), we have focused on two aspects of the process to see what is characteristic to them: the creation of visualizations (question 2a) and exploratory results (question 2b).

7.2.1 Creation of visualizations

The complexity of inter-organizational networks offers many approaches to their analysis and visualization. They can be thought of as social networks where nodes are organizations rather than individuals. This approach will allow standard network analysis. However, these organizations contain individuals, who may be extremely important to the network as a whole. In our graph-based visualization, we presented nodes of several different types at the same time: individuals, keywords and organizations. The types relevant to a particular network vary, but it might be useful to think of the network in terms of many different types of components.

In our study, the most interesting features of the network turned out to be the definition of a relationship. As discussed in the beginning of this thesis,

relationships in inter-organizational networks can mean many things, such as cooperation, competition or dependency. In our case, we focused explicitly on cooperation and tried different ways to measure this, such as project count, man months and money. Another point of interest was details about individual organizations and relationships. While it is likely that given data about the whole network, the structure of the network could also reveal very interesting things, it seems that at least in some contexts, the practical value of understanding networks is still grounded in such details, which should be accessible in visualizations.

We also showed that the visualization of networks does not need necessarily to be limited to graphs and matrices. Our time series, based on a simple concept, was as well received by our domain experts as our graph-based visualization. What they were most interested in was not the specific presentation, but what kind of themes a particular visualization was able to expose. For the time series, it was naturally evolution over time, while our graph was focused on connections between different kinds of entities.

Our visualizations barely took advantage of the analytical tools offered by network theory. No shortest paths or average degrees were explicitly computed or their significance even contemplated. Still, significant exploratory results were gained. Just because one is dealing with networks does not mean that such approach is necessary by default. Nevertheless, one could also ask if even better results would be possible if network theory was also considered already in the exploratory phase.

7.2.2 Exploratory results

Three main types of insight arose during our case study: exploratory themes, such as key people within organizations in the network, specific questions, and simple facts. Of these, general themes were most numerous, although the more concrete findings became more frequent as the exploration matured. In addition to these themes, the exploration revealed the characteristics of the currently available data, which was considered an important result in itself.

Themes related to inter-organizational networks build an overview of the network from a particular perspective and also serve as a basis for further analysis. They are particularly needed when the data set is explored for the first time. Questions address specific facts about the data and may reveal interesting patterns. They may drive the development of visualizations, or serve as a basis for some other kind of analysis. Facts are very specific and are closest to the practical reality of the networks, such as the amount of value created by a particular relationship.

Could these kind of results be produced by other means, such as data

mining? This is quite possible. However, collaborative interpretation with domain experts would have been hard to implement without a compelling visual presentation of the data. It is also likely that while data mining or other non-visual analysis would have produced more exact results, such as specific models, based on the subset of data that was workable with those techniques, it would have missed other themes and failed to provide such an easily communicable overview of the data available.

Chapter 8

Conclusions

This study has provided an overview of the exploratory visualization process in literature, and included a case study of the process in a particular context. By focusing on a particular, but growingly important data type we were able to conduct a qualitative case study that provided insight into the practical reality of the process that has thus far gained little attention in research. In the following, we reiterate the most important findings in this study, consider its limitations, and discuss future work.

8.1 Contribution

The goal of this study was to 1) enumerate and describe the steps of the exploratory visualization process, and 2) examine the exploratory visualization of inter-organizational networks in particular. We constructed a process model of the exploratory visualization process based on literature from several related fields, and presented a case study describing the exploration of inter-organizational networks in a particular context. From these, we drew design implications for the development and evaluation of tools meant to support this process, or as a basis for designing and managing such expeditions. We hope that these results will not only help the development of better software, but also encourage the adoption of visualization as a practical tool outside of the scientific community.

The exploratory visualization process was distilled into seven steps: 1) problem definition, 2) data discovery and selection, 3) data preprocessing, 4) creation of visualization(s), 5) interpretation, 6) evaluation, and 7) presentation. These are not necessarily consecutive, and iteration between steps and backtracking to previous ones is expected. This model however shows that exploratory visualization involves much more than just creating visualiza-

tions (whether this is taken to mean using an existing visualization system or building one).

Particularly interesting for developers of visualizations is considering where the data of a visualization comes from and how to support heterogeneous data sets, data migration, and metadata, recognizing interpretation as a separate task from the construction of visualizations, and supporting the presentation or deployment of exploratory results in some form or another. It is also useful to keep in mind the various stakeholders involved in exploratory visualization: besides the visualizer themselves, possible collaborators and domain experts, data gatekeepers, and the consumers of the exploratory results.

In fact, the process of exploratory visualization is not unlike that of data mining. The steps of the process are similar, understanding the domain of the data is important, and the process is iterative. However, there are some key differences. Visualization makes it easier to share and collaborate, interpretation can happen on several levels (user interaction, iterative creation of visualizations, and discussion), and the exploratory results are likely to be different in nature (themes and data-specific questions and facts rather than rigorous models or patterns). Visualization can also better deal with heterogeneous and noisy data than pure data mining.

We developed two new visualizations in our case study that were very enthusiastically received. From them, we can draw a few observations regarding the visualization of inter-organizational networks. First, they can effectively be visualized by other means than graphs. Simple plots depicting a relevant dimension of the network, such as time, can be just as effective. For graphs, different types of nodes depicting different elements of the network, such as individuals, organizations, or even keywords, can be shown simultaneously to enable the detection of complex relationships. Network theory is not the only way to approach networks – in our case, details about individual organizations and their relationships turned out to be more relevant. Finally, when depicting complex inter-organizational networks, the definition and measure of a relationship between two organizations must be considered according to the perspective of the analysis.

8.2 Limitations

We have tried to construct a model that describes exploratory visualization in its most general sense. However, the field is wide and includes very different practices. Where applicable, we have focused specifically on early exploration within organizations that have access to some data but may not have experience in visualizing it. The overall structure of the model is likely

to be generalizable to any exploratory visualization, but the specifics of the steps may vary according to context.

The case study has similar limitations. Most notably we only considered cooperation networks as a special case of inter-organizational networks. Our setting was also such that the visualizer is not very familiar with the domain of the data, unlike in many cases where analysts are experienced with data from their own domain, so the roles of the visualizer and the domain expert were inherently separate. This setting applies best to the development of new visualization systems as well as organizations where visualization or data mining are not everyday practices.

What may be controversial about the approach in this study is the use of "second-class" data which is incomplete, heterogeneous and not rigorously screened for consistency. We argue that part of the appeal of visualization is its ability to tap into such data, and make the quality of the data transparent to those familiar with its domain. However, such fuzzy approach may not always be applicable.

8.3 Future work

Although our case study has finished, the visualizations developed as a part of it will continue to be used by domain experts to explore Aalto's cooperation network. Evaluation of their value as practical tools should be conducted in this environment. We are planning to do this by logging their use and analyzing these logs.

The model presented here also opens several interesting questions for further research into exploratory visualization. Its applicability in other contexts and with other kind of data should be investigated. Our conclusions of the relationship of information visualization and data mining could be verified by a comparative study, allowing further examination the strengths and similarities of the two approaches to data. Finally, and most importantly, methodology for the least well-defined step of the process, the evaluation of exploratory results, should be developed.

The analysis of inter-organizational networks is likely to become more relevant and interesting as data analysis tools and network theory both continue to develop. The visualization of these networks is still in its infancy. The different perspectives of ego-centric and network-level can be investigated further to understand when they are relevant, especially in terms of organizations trying to understand the network they are themselves part of. Methodology, both in terms of visualization techniques and organizational processes for grasping such networks can be further developed.

Bibliography

Robert Amar and John Stasko. Best paper: A knowledge task-based framework for design and evaluation of information visualizations. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 143–150. IEEE, 2004.

Bruce R Barringer and Jeffrey S Harrison. Walking a tightrope: creating value through interorganizational relationships. *Journal of management*, 26(3):367–403, 2000.

Rahul C Basole. Visualization of interfirm relations in a converging mobile ecosystem. *Journal of Information Technology*, 24(2):144–159, 2009.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009. URL `http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154`.

Sean Bechhofer, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Philip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, Matthew Gamble, Danius Michaelides, Stuart Owen, David Newman, Shoaib Sufi, and Carole Goble. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2):599 – 611, 2013. ISSN 0167-739X. doi: 10.1016/j.future.2011.08.004. URL `http://www.sciencedirect.com/science/article/pii/S0167739X11001439`.

Guido Caldarelli and Alessandro Vespignani. *Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science (Complex Systems and Interdisciplinary Science)*. World Scientific Publishing Company, 2007.

Steven P Callahan, Juliana Freire, Emanuele Santos, Carlos E Scheidegger, Cláudio T Silva, and Huy T Vo. Vistrails: visualization meets data management. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 745–747. ACM, 2006.

Stuart K Card, Bongwon Sun, Bryan A Pendleton, Jeffrey Heer, and John W Bodnar. Time tree: Exploring time changing hierarchies. In *Visual Analytics Science And Technology, 2006 IEEE Symposium On*, pages 3–10. IEEE, 2006.

Sheelagh Carpendale. Evaluating information visualizations. In *Information Visualization*, pages 19–45. Springer, 2008.

Chaomei Chen and Mary P. Czerwinski. Empirical evaluation of information visualizations: an introduction. *International Journal of Human-Computer Studies*, 53(5):631 – 635, 2000. ISSN 1071-5819. doi: 10.1006/ijhc.2000.0421. URL `http://www.sciencedirect.com/science/article/pii/S107158190090421X`.

Bender S. Demoll and D. Mcfarland. The Art and Science of Dynamic Network Visualization. *JoSS: Journal of Social Structure*, Volume 7, 2005. URL `http://www.cmu.edu/joss/content/articles/volume7/deMollMcFarland/`.

D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010. ISBN 9780521195331.

Geoffrey Ellis and Alan Dix. An explorative analysis of user evaluation studies in information visualisation. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, BELIV '06, pages 1–7, New York, NY, USA, 2006. ACM. ISBN 1-59593-562-2. doi: 10.1145/1168149.1168152. URL `http://doi.acm.org/10.1145/1168149.1168152`.

Elsevier. Scopus, 2013. URL `http://www.scopus.com/`. Accessed 2013-04-23.

Linton C. Freeman. Visualizing social networks. *Journal of Social Structure*, 1(1), 2000.

Ben Fry and Casey Reas. Processing, 2004-2013. URL `http://processing.org/`. Version 20b7.

Mohammad Ghoniem, J-D Fekete, and Philippe Castagliola. A comparison of the readability of graphs using node-link and matrix-based representations. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 17–24. IEEE, 2004.

Google. Google charts. URL `https://developers.google.com/chart/`. Accessed 2013-04-17.

J. Han, M. Kamber, and J. Pei. *Data Mining, Second Edition: Concepts and Techniques*. Data Mining, the Morgan Kaufmann Ser. in Data Management Systems Series. Elsevier Science, 2006. ISBN 9780080475585.

Tom Heath and Christian Bizer. Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1):1–136, 2011.

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

Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 203–212. ACM, 2010.

Martin Hilbert and Priscila López. The world's technological capacity to store, communicate, and compute information. *Science*, 332(6025):60–65, 2011.

TJ Jankun-Kelly, Kwan-Liu Ma, and Michael Gertz. A model and framework for visualization exploration. *Visualization and Computer Graphics, IEEE Transactions on*, 13(2):357–369, 2007.

Sean Kandel, Andreas Paepcke, Joseph M Hellerstein, and Jeffrey Heer. Enterprise data analysis and visualization: An interview study. *Proc. IEEE Visual Analytics Science & Techology*, 2012.

Hyunmo Kang, Catherine Plaisant, Bongshin Lee, and Benjamin B Bederson. Netlens: iterative exploration of content-actor network data. *Information Visualization*, 6(1):18–31, 2007.

M. Kantardzic. *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons, 2011. ISBN 9780470890455.

D.A. Keim. Information visualization and visual data mining. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):1 –8, Jan/Mar 2002. ISSN 1077-2626. doi: 10.1109/2945.981847.

Martin Krzywinski, Jacqueline Schein, İnanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics, 2009.

Martin Krzywinski, Inanc Birol, Steven JM Jones, and Marco A Marra. Hive plots—rational approach to visualizing networks. *Briefings in Bioinformatics*, 13(5):627–644, 2012.

Bongshin Lee, Cynthia S Parr, Catherine Plaisant, Benjamin B Bederson, Vladislav D Veksler, Wayne D Gray, and Christopher Kotfila. Treeplus: Interactive exploration of networks with enhanced tree layouts. *Visualization and Computer Graphics, IEEE Transactions on*, 12(6):1414–1426, 2006.

MATLAB. *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts, 2010.

Dan McFarland and Skye Bender-deMoll. Sonia - social network image animator. URL `http://www.stanford.edu/group/sonia/index.html`. Accessed April 18, 2013.

Glenn J Myatt and Wayne P Johnson. *Making sense of data II: a practical guide to data visualization, advanced data mining methods, and applications*, volume 2. Wiley, 2009.

T. Neirynck and K. Borner. Representing, analyzing, and visualizing scholarly data in support of research management. In *Information Visualization, 2007. IV '07. 11th International Conference*, pages 124 –129, July 2007. doi: 10.1109/IV.2007.94.

Petra Neumann, Anthony Tang, and Sheelagh Carpendale. A framework for visual information analysis. Technical report, Technical Report 2007-87123, University of Calgary, Calgary, AB, Canada, 2007.

Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

Mark EJ Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5200–5205, 2004.

Martin Šperka and Peter Kapec. Interactive visualization of abstract data. *Science & Military*, 5(1):84–90, 2010.

Catherine Plaisant. The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced visual interfaces*, pages 109–116. ACM, 2004.

Zachary Pousman, John T Stasko, and Michael Mateas. Casual information visualization: Depictions of data in everyday life. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1145–1152, 2007.

Keith G Provan, Amy Fish, and Joerg Sydow. Interorganizational networks at the network level: A review of the empirical literature on whole networks. *Journal of management*, 33(3):479–516, 2007.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org`. Accessed 2013-04-17.

Thompson Reuters. Web of science, 2013. URL `http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/`. Accessed 2013-04-17.

N. Rubens, M. Russell, R. Perez, J. Huhtamaki, K. Still, D. Kaplan, and T. Okamoto. Alumni network analysis. In *Global Engineering Education Conference (EDUCON), 2011 IEEE*, pages 606 –611, April 2011. doi: 10.1109/EDUCON.2011.5773200.

P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *Visualization and Computer Graphics, IEEE Transactions on*, 11(4):443–456, 2005. ISSN 1077-2626. doi: 10.1109/TVCG.2005.53.

Michael Sedlmair, Petra Isenberg, Dominikus Baur, and Andreas Butz. Information visualization evaluation in large companies: Challenges, experiences and recommendations. *Information Visualization*, 10(3):248–266, 2011.

B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.

R. Spence. *Information Visualization: Design for Interaction*. Pearson/Prentice Hall, 2007. ISBN 9780132065504.

Robert Spence. A framework for navigation. *International Journal of Human-Computer Studies*, 51(5):919–945, 1999.

Rebecca R Springmeyer, Meera M Blattner, and Nelson L Max. A characterization of the scientific data analysis process. In *Proceedings of the 3rd conference on Visualization'92*, pages 235–242. IEEE Computer Society Press, 1992.

Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1983.

Edward R. Tufte. *Envisioning Information*. Graphics Press, 1990.

Edward R. Tufte. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, 1997.

Hans van der Heijden. *Designing management information systems*. Oxford University Press, USA, 2009.

Aaron Wildavsky. Information as an organizational problem. *Journal of Management Studies*, 20(1):29–40, 1983.

Hui Xu. A regional university-industry cooperation research based on patent data analysis. *Asian Social Science*, 6(11):p88, 2010.

Ji Soo Yi, Youn-ah Kang, John T Stasko, and Julie A Jacko. Understanding and characterizing insights: how do people gain insights using information visualization? In *Proceedings of the 2008 conference on Beyond time and errors: novel evaluation methods for Information Visualization*, page 4. ACM, 2008.