Ostinato: The exploration-automation cycle of user-centric, process-automated data-driven visual network analytics

Jukka Huhtamäki
Tampere University of Technology

Martha G. Russell
mediaX at Stanford University

Neil Rubens
University of Electro-Communications

Kaisa Still
VTT Technical Research Centre of Finland

Abstract Network analysis is valuable method for investigating and mapping the social structure driving phenomena and sharing the findings with others. The interactive visual analytics approach transforms data into views that allow the visual exploration of the structures and processes of networks represented by data, therefore increasing the transparency of editorial processes on social media as well as networked structures in innovation ecosystems and other phenomena. Although existing tools have opened many new exploratory opportunities, new tools in development promise investigators even greater freedom to interact with the data, refine and analyze the data, and explore alternative explanations for networked processes. This chapter presents the Ostinato Model – an iterative, user-centric, process-automated model for data-driven visual network analytics. The Ostinato Model simultaneously supports the automation of the process and enables interactive and transparent exploration. The model has two phases, Data Collection and Refinement and Network Creation and Analysis. The Data Collection and Refinement phase is further divided into Entity Index Creation, Web/API Crawling, Scraping, and Data Aggregation. The Network Construction and Analysis phase is composed of Filtering in Entities, Node and Edge Creation, Metrics Calculation, Node and Edge Filtering, Entity Index Refinement, Layout Processing and Visual Properties Configuration. A cycle of exploration and automation characterizes the model and is embedded in each phase.

Keywords: data-driven, network analysis, visual analytics, process model, social media analytics, method development

1. Introduction

This chapter introduces the Ostinato Model, an exploration-automation cycle for a user-centric, process-automated, data-driven visual network analytics.

In terms of increasing the transparency of editorial processes on social media, this chapter contributes to the general theme of the book and particularly its second volume at hand in three levels. First, network analysis is a key approach in supporting explorative studies on the patterns and structures in between actors creating, curating, refining, and distributing social media content and in estimating the authority and trust these actors have, therefore allowing for increasing the transparency of the editorial structure of Wikipedia co-authors, discussion and dissemination structures on Twitter and other social media. These structures can be modeled, represented, analyzed and visualized as networks to support the investigations and exploration. Second, the presented data-driven approach allows extending these investigations of patterns

and structures within and in between groups of actors can be extended beyond the boundaries of individual social media and to happen over long periods of time. Third, actors with different sets of skills from means to crawl online sources for data to domain knowledge allowing deep sensemaking can all fully engage into the different phases of the investigative process.

These contributions allow the use of visual representations of the structures behind various social media phenomena to improve social interaction, estimations of trust and credibility on social media. With the data-driven approach, the investigators of social media phenomena and patterns of social interaction, trust and credibility are able to move fast in the beginning of the process. As the ways of visualizing and investigating a particular phenomena matures, the investigators may continue to follow the phenomena with the support of close to real-time dashboards adding transparency and supporting e.g. longitudinal investigations. The option for automating the process also supports developing these investigative tools toward end-user products for avid social media content authors and users.

In music, the word "ostinato" refers to both a repeating musical pattern as well as a composition that contains a repeating musical pattern. Like the repeated rhythms and melodies in Ravel's *Bolero* – small innovations are explored with each iteration, and some are incorporated into the melodic narrative – we apply the musical concept of "ostinato" to a cycle of user-centric exploration and automation that builds transparency of authorship for evidence-based decision making.

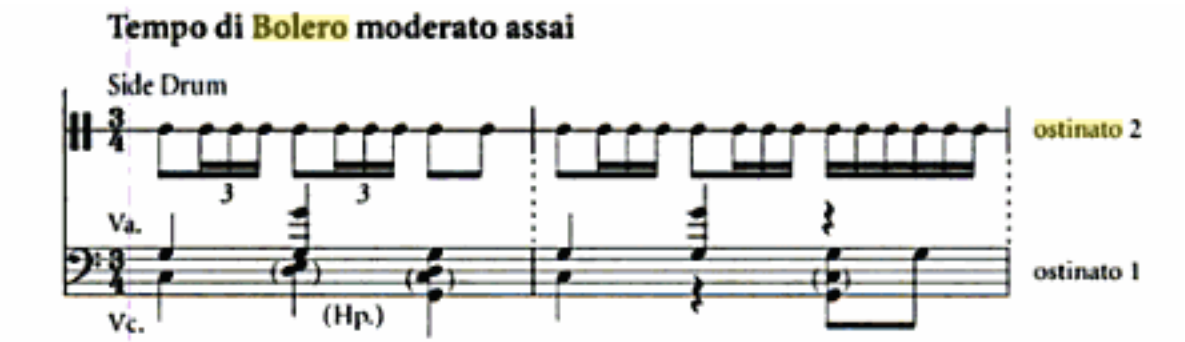

Figure 1.1 Ostinato patterns from Bolero's *Ravel* (Dawer 2000*)*

Here, data-driven means that the analysis process relies on data, is automated and conducted in a computational manner, and visual network analytics refers to taking a visual analytics (Thomas and Cook 2006, Heer and Shneiderman 2012) approach to network analysis. Additional data can augment the dataset selected for analysis through an automated software process. Established analytical procedures can be automated, yet new conditions for analysis-based insights can be introduced and refined incrementally with continuous computational iterations.

In this implementation of the Ostinato Model, the phenomena under investigation are modeled as a network, and highly interactive visualization tools are used to conduct the investigative process. Network analysis introduces a relationship approach to investigating the structure of many kinds of phenomena. Network analysis allows for exploratory analysis of the social roles

of network actors and the phenomena of relationships, as well as for quantifying the structural properties of networks.

A key aspect of the Ostinato Model is the focal point of the user – here, the investigator of particular network-driven phenomena – in the investigative process. This answers to the call for data scientists[1], somewhat mythical multi-skilled individuals that are capable of individually running the whole investigative process from collecting data to analysis to deep sensemaking in domain of interest, by allowing both experts of the domain under investigation, developers of the technical process as well as e.g. quantitative analysis specialists to possess equal means to take a proactive role in the investigative process. Moreover, the Ostinato Model defines an overall structure for the data-driven investigative process that supports the coordination between the individual phases of the process and therefore allows all the members of the investigative team to contribute to the implementation of different phases of analysis.

Visual network analytics allows the emergence of insights on the structure and dynamics of innovation ecosystems, social media platforms and other networked phenomena. Existing research on networks shows that network analysis has a good fit for explorative analysis of (eco)systems: much is already known about structure in networks (Granovetter 1973, Barabási and Bonabeau 2003), the roles of individual actors in the network (Hansen et al. 2011), the drivers of network evolution (Giuliani and Bell 2008) as well as the latent structures and dynamics behind the diffusion of information through networks (Leskovec et al. 2009), network control (Liu et al. 2011) and virality (Shakarian et al. 2013, Weng et al. 2013). Transforming those insights into action requires communicating the insights to constituents of change (Russell et al. 2011, Still et al, 2014). Visual network analysis is a promising method for investigating social configurations and for interactively communicating their findings to others (cf. Freeman 2009).

Data-driven visual network analytics leverages computation to analyze potentially very large datasets in order to identify the patterns driving complex phenomena. Moreno (1953), Freeman (2000, 2009), Hansen et al. (2009, 2011), Russell et al. (2011), Still et al. (2014), Basole et al. (2012), Ritala and Hallikas (2011), and Ritala and Huizingh (2014) give examples of using a network approach to investigate complex phenomena that are driven by sets of interconnected actors. The investigations of such phenomena are further complicated because data about these actors frequently come from multiple and diverse data sources, some of which are not developed for computational use. Especially in cases involving data that are heterogeneous by nature, an iterative, incremental analysis process is sometimes necessary (Telea 2008). Analysis of complex phenomena often involves multiple pathways to actionable recommendations, and assumptions underlying decisions may change over time.

We agree with Freeman (2000) that integrated tools that can be used to collect, manage and visualize the SNA data are key in supporting network investigations. The tradeoff between usability and automation sometimes creates a barrier for new entrants into data-driven visual network analysis (Hansen et al. 2009). In order to provide a low barrier approach to using

[1] Ideally, a data scientist is a hacker, scientist, quantitative analyst, trusted adviser and business (domain) expert, all in one person (cf. Davenport 2014).

network analysis to study complex phenomena, we prioritize usability over process automation when possible.

However, a gap exists between the vision and the practice. Manually operated processes used by individual investigators or small investigative teams rely on ready-made tools that are operated through graphical user interfaces. Using these stand-alone tools is very straightforward. The available data sources and analysis and visualization functionalities are, however, somewhat limited. The full-stack, programming-centric processes, in which massive sets of data are mined with tools that are developed and operated by experts, are generally run in complex cloud-based environments. We are aware that several process models, with different levels of abstraction, exist to structure data-driven, visualization-centric investigations; a selection of these models will be covered as part of the description of previous work in the next section.

Many of the existing models are either very general or focus on particular parts of the process. A data-driven visual network analytics approach requires drawing from a number of process models. Using parallel data sources is often not considered in the process models. Moreover, network analysis introduces specific requirements to the process, importantly including the possibility to calculate node metrics as additional data quantifying the different structural roles of the nodes.

Drawing from our experience in running multiple case studies in the context of explorative innovation ecosystem analysis, we take a design science research (Hevner et al. 2004) approach to describe a process model for data-driven visual network analytics. In this book, our chapter contributes to the body of knowledge on computational frameworks, tools and algorithms for supporting transparent authorship in social media knowledge markets by defining an interactive and iterative process model for data-driven visual network analytics to explore relationships in ecosystems. Our process model takes into account requirements stemming from a call for transparent authorship in social media knowledge markets and builds on existing models for data-driven analytics and sensemaking. It is designed to support iterative and incremental investigative processes, as well as to automatically update a visualization dashboard revealing the dynamics and evolving network structure of a phenomenon under investigation.

In this chapter, we introduce the Ostinato Model – an iterative, user-centric, process-automated model for data-driven visual network analytics that relies on data from multiple social media and other online sources. In music, an ostinato is a repeating, persistent rhythmic pattern, melody or motif in a consistent musical voice. Similarly, ostinato patterns emerge when conducting data-driven studies on innovation ecosystems and related phenomena.

The rest of the chapter is organized as follows. In second section, we review previous work on which this Ostinato process model is based. The third section introduces the research methodology and a selection of cases we have used to develop the model. The fourth section describes the requirements for the process as well as the different steps that constitute the Ostinato process model. In the fifth section, we discuss how this model satisfies these criteria and adapts to the exploration – automation cycle. The sixth section concludes the chapter and describes key implications and ideas for future work.

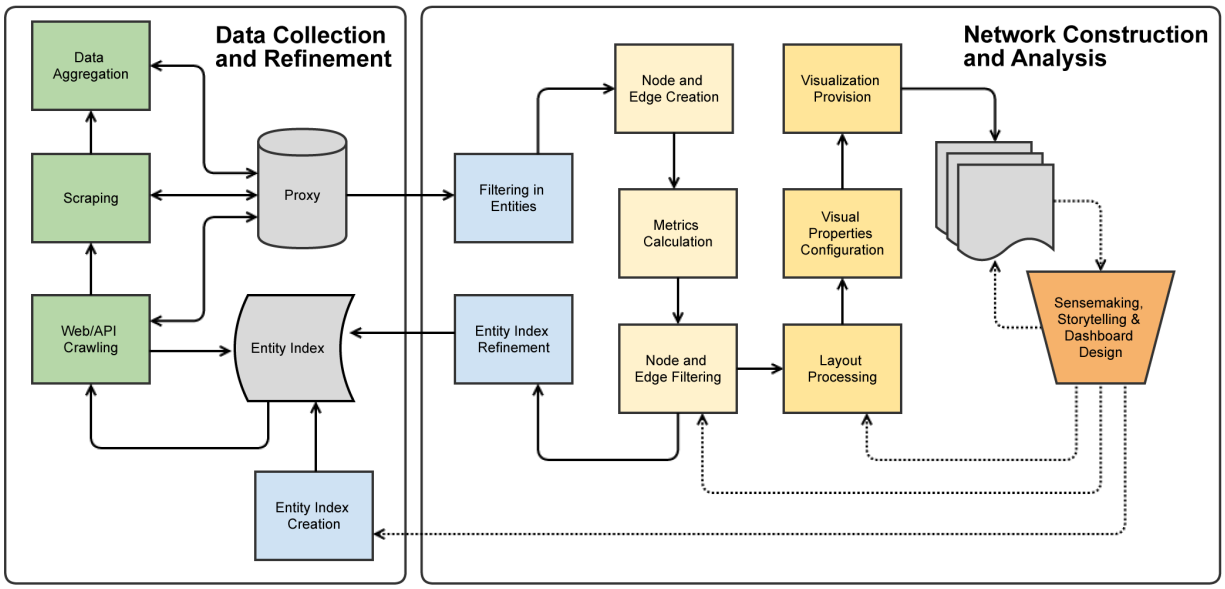

Figure 1.2 Ostinato Model – user-centric data-driven process model for visual network analytics

2. Previous work

Our approach into data-driven visual network analytics builds on a number of bodies of knowledge, including traditional SNA (Wasserman and Faust 1994), information visualization (Card et al. 1999), data-driven visualization pipelines (Nykänen et al. 2008), interactive network analysis (Hansen et al. 2009), visual analytics (Thomas and Cook 2006), sensemaking (Pirolli and Card 2005), interactive visualization (Heer and Shneiderman 2012) and scientific visualization (Telea 2008). All these fields offer models and approaches, and additionally they pose key requirements to be considered when developing next-generation analytics tools for very large networks. The objective to conduct (and publish) research in a reproducible way (Peng 2009, Ghosh 2013) contributes to the quality of the process and also introduces additional requirements.

Traditional SNA (Wasserman and Faust 1994) introduces a set of node and network level metrics that can be used to describe the structural properties of networks and to quantify the various social roles of network actors. To support the use of network analysis, Hansen et al. (2009) build on top of the sensemaking model to present the Network Analysis and Visualization (NAV) process model to support novices that enter network analysis. The NAV process starts with defining the goals for the analysis and continues through data collection and structuring, after which data are interpreted through multiple loops of network visualization and SNA metrics calculation. Finally, the insights and conclusions are formatted and summarized, then disseminated through a report. Seeking low-barrier entry, the authors introduce NodeXL, an Excel-based toolset for SNA, to conduct the analysis. Among others, Hansen et al. (2011) define ways to apply these metrics in investigating phenomena taking place in social media.

The information visualization reference model (Card et al. 1999) presents a four-step process that can be used as a blueprint for implementing data-driven visualization processes. Raw data is (1) first collected and then (2) refined into data tables to allow straightforward processing.

Data tables are then (3) transformed into a portfolio of visual representations from which various concrete views are (4) served to the visualization user for sensemaking. Importantly, the reference model suggests that best practice is when the user can interact with all steps of the process.

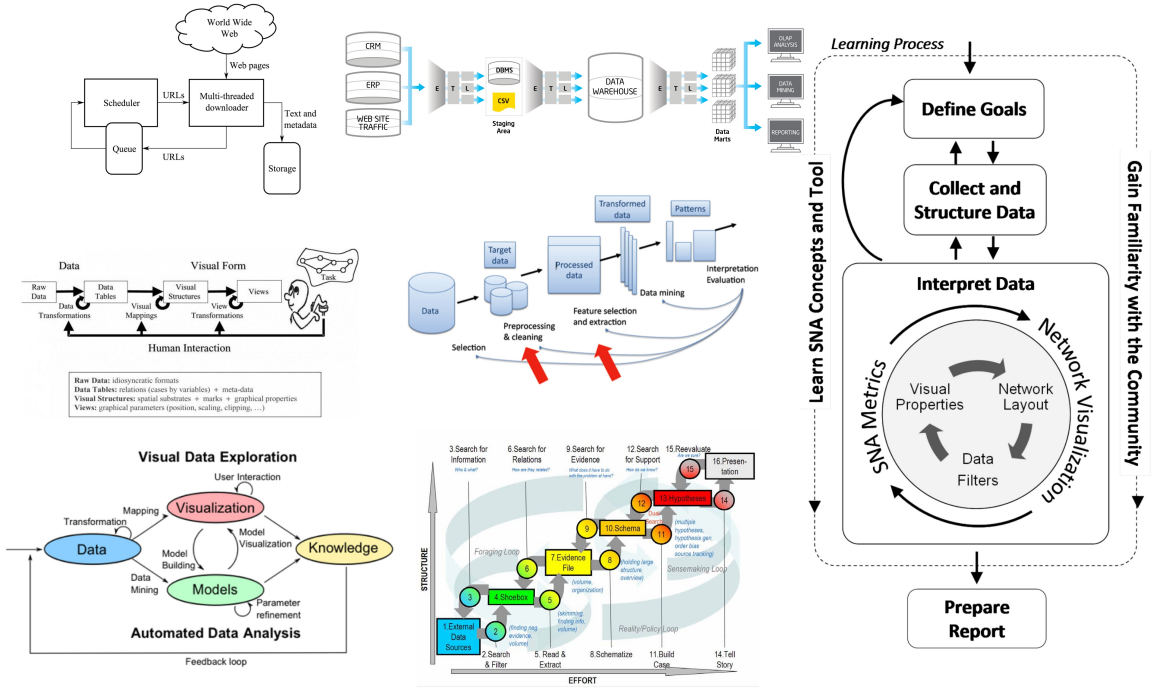

Figure 2.1 Process models related to data-driven visual network analytics. The six small diagrams, from top-left: Web Crawling (Wikipedia 2014), Extract-Transform-Load (Intel 2013), Information Visualization Reference Model (Card et al. 1999), Knowledge Extraction from Databases (Indarto 2013), Visual Analytics (Keim et al. 2010), Sensemaking (Pirolli and Card 2005). On the right: Network Analysis and Visualization (NAV) model (Hansen et al. 2009).

Component-based data-processing pipelines, a technical application of the information visualization reference model, introduce a viable approach for developing reusable pieces of software to support the automation of processes related to social network analysis across application domains (Nykänen et al. 2008, Huhtamäki et al. 2010). To support investigating the social structure among wiki co-creators, Huhtamäki et al. (2010) present a set of components and a process model to orchestrate the use of the components. A key benefit of the component-based approach presented by Nykänen et al. (2008) is the possibility to integrate existing software tools implemented in different technologies into the data-processing pipeline, given that they can be operated from the command line. The main restriction of the approach is the need to implement the automation through scripting, i.e. writing program code that describes rules for a particular functionality.

The general sensemaking model by Pirolli and Card (2005) divides the sensemaking process into two loops, the foraging loop and the sensemaking loop. To simplify, data is first collected and refined and then transformed into various visualizations and other representations that support the sensemaking. The process is iterated as many times as required. Similarly, the process of visual analytics “typically progresses in an iterative process of view creation, exploration, and refinement” (Heer and Shneiderman 2012).

The sensemaking step can be applied in different ways - from purely manual processes in which humans interact with various user interfaces to conduct the analysis to automated information systems in which data are collected and processed in runtime. Sensemaking also includes the process of visual analytics (Thomas and Cook 2006) that, by default, relies on the availability of software and tools supporting the users. Heer and Shneiderman (2012) give an insightful overview to the specific functionalities that users should be able to operate: (1) specify data and views; (2) manipulate views; and (3) process and provenance their findings.

Peng (2009) builds his definition of reproducible research on three categories: a piece of research is fully reproducible if both the data and code used to are available and, moreover, if the code is executable by anyone. As Ghosh (2013) shows, reproducibility can be approached at many different levels, from policy to detailed technological solutions.

3. Methodology

In this section, we briefly describe the context in which the data-driven network analytics takes place. This illustrates the explanatory power and novelty introduced by the Ostinato Model to network analysis workflows. Further, we discuss the use of Design Science Research (Vaishnavi and Kuechler 2007) as a method that we apply in our venture to develop the Ostinato Model in a way that is both credible in terms of scientific theory . We also refer to a selection of case studies we have used to develop and validate the process model presented in this chapter. This, on the other hand, shows that the Ostinato Model is a general approach to conducting data-driven network analysis investigations and that it has already been extensively applied in a series of real life experiments and investigations.

Addressing innovation ecosystems as networks allows scholars and practitioners to study their complexity, providing a means for mapping, monitoring and managing the ecosystem components. To do this, we have taken a data-driven network analysis approach to study innovation ecosystems in regional, metropolitan, national and international level as well as e.g. in the context of programmatic activities supporting innovation and growth. We have followed a design science research approach that is based on iteration through construction of network visualizations as artifacts. We have used a number of different datasets in these studies, including social media, socially constructed data available online, and proprietary sets of data represented as spreadsheets and other formats.

3.1. Context

The research that led to the development of this process model for data-driven network analysis began in the context of studying complex networks of relationships in innovation ecosystems. Russell et al. (2011) use the concept of innovation ecosystem to refer to the inter-organizational, political, economic, environmental, and technological systems through which a milieu conducive to business growth is catalyzed, sustained, and supported. A dynamic innovation ecosystem is characterized by a continual realignment of synergistic relationships that promote growth of the system. (Russell et al. 2015)

Ecosystems are a complex phenomenon, with multiple entities connected through multiple level relationships, as well as multiple stakeholder perspectives into those relationships. Ecosystems that promote innovation have become a quest for companies, cities, regions and countries. It is

agreed that “relationships shape the behavior and outcome of all stakeholders as well as the system-level effects” (Hwang and Horowitt 2012), and that it is through the relationships of individuals within and across organizations in an ecosystem that knowledge transfer, technology dissemination and organizational change are accomplished (Russell et al. 2015). Program managers and policy analysts in charge of transforming innovation ecosystems seek to define and describe innovation ecosystems in order to set goals, determine interventions and evaluate change, and visualizing the innovation ecosystem has proven instrumental to strategy setting and decision-making (Still et al. 2014). By making the roles and relationships explicit, both numbers and visualizations can be used to support the creation and management of innovation ecosystems. By tracking the provenance of data and authorship of analytical refinements, the collaborative exploration gains transparency.

To manage as well as to create innovation ecosystems, network orchestration has been encouraged (Ritala 2009, Paquin and Howard-Grenville 2013, Still et al. 2014). A data-driven process for understanding roles allows for interactive discovery of the innovation ecosystem. Multiple perspectives can be invited and exchanged in the process of developing and orchestrating transformation programs. With subsequent automation of data updates and tracking analyses, the assumptions and contingencies underlying decisions can be monitored for changes that would impact policy and program directions.

3.2. Design science research

In this research, we take a Design Science Research (DSR) approach to describe a process model for data-driven visual network analytics applicable in replicable investigations of innovation ecosystems as well as other domains in which network structures over time are of interest. DSR is a research method that allows “learning and investigation through artifact construction” (Vaishnavi and Kuechler 2007, p. 187). “Whereas natural sciences and social sciences try to understand reality, design science attempts to create things that serve human purposes” (Simon 1969). The rationale for DSR hails from the importance of the practical utility of research (Peffers et al. 2007). Design science research aims to build a bridge between information system (IS) research and its practical application by producing results that have real-life relevance. “Design science [...] creates and evaluates IT artifacts intended to solve identified organizational problems” (Hevner et al. 2004).

The General Design Cycle (GDC) is a key part of any DSR process (Vaishnavi and Kuechler 2007). The process begins from the awareness of the problem and continues to one or more suggestions for solution. Next, an implementation of the plan is developed and evaluated, and finally, the process is concluded and the results shared; in the case of a scientific process, they are published. In each of these steps, new knowledge is both created and fed back to previous phases. The phases are repeated in an iterative fashion until a satisfactory end result (one that has practical utility) is achieved.

Readers with experience in software development will notice a straightforward connection between general design cycle and agile software development (see e.g. Schwaber and Beedle 2001). Apart from the intent to publish the results, both design and development processes move forward in an iterative and incremental fashion and are guided by feedback collected from

the users and other stakeholders of the developed software or other artifact, here the process model.

To develop the Ostinato Model for data-driven visual analytics presented in this chapter, we effectively applied and repeated the General Design Cycle. To evaluate the process model for added credibility of the presented results, we applied the Experimentation Pattern defined by Vaishnavi and Kuechler (2007), more specifically the case-based prototype development pattern on which the prototype is developed in an incremental, iterative manner over a number of cases, leading to deep knowledge of the problem and the proposed solution.

3.4 Example cases

The Ostinato Model has been developed over a number of cases in which a variety of innovation ecosystems have been investigated in collaboration with their stakeholders using various sets of data sources (Rubens et al. 2011, Basole et al. 2012; Russell et al. 2015; Jussila et al. 2014). Table 1.1 describes core cases in which the automation-exploration cycle was implemented, using structured and semi-structured data sources, involving stakeholders in the exploration process as well as the sensemaking of key visualizations and other outputs for each case.

Table 1.1 Illustrative cases for developing the exploration-automation cycle and the process model for data-driven visual network analytics

Case	Data	Co-creators/case shareholders	Visualizations/Outputs
Demola (Huhtamäki et al. 2013)	Proprietary data on Demola projects, the companies that initiated the project and university affiliations of project members (university students)	Demola leaders and operators and the investigative team	The animation of the evolution of Demola project sphere including projects, the affiliations of project team members and companies. Multimode networks on (1) projects and affiliated actors and (2) projects and their key competences
Tekes Young Innovative Companies (Huhtamäki et al. 2012)	Innovation Ecosystem Network Dataset on growth companies, Twitter data on Tekes Young Innovative Companies (YIC) and their followers	Policy makers at Tekes - the Finnish Funding Agency for Innovation and the investigative team	One and two-step networks of the companies part of Tekes YIC program and their affiliations to investors and key individuals
Finnish Innovation Ecosystem (Still et al. 2013)	Three separate datasets: 1) Thomson Reuters SDC for deals and alliances and IEN Dataset for 2)	Finnish national-level policy makers and the investigative	Network visualizations and metrics about companies having their main office in Finland and their first-step connections to other companies, investors and

	Executives and Finance and 3) Startups and Angels	team	key individuals
Network orchestration for EIT ICT Labs (Still et al. 2014)	IEN Dataset for Executives and Finance	EIT ICT Labs representatives and the investigative team	Network visualizations of companies having their main office in one of the EIT ICT Labs co-location centers and their first-step connections to investors and individuals as well as to other companies through investments and acquisitions

4. Ostinato model

This section presents summary of the results of our research. First, we describe the requirements for the data-driven visual network analytics process; these requirements stem from existing process models and are augmented through results that emerged in case studies on which we applied the method. Second, as the core contribution of this chapter, we describe the process model for the exploration – automation cycle of data-driven visual analytics, the Ostinato Model.

4.1 Process requirements

Developed through several rounds of iterations following the General Design Cycle, the core guidelines and requirements for the data-driven visual network analytics process model include the following: continuous data collection; exploration; transparency; loose coupling; reproducibility; automation; enabling manual steps; low entry barrier; and interoperability. Each is described.

Continuous data collection. When collecting data from social media, persistent processes are often needed, particularly when the investigators want to capture both the structure and dynamics of a phenomenon. Twitter, for example, currently provides only limited access to its historical data, and even then data on followers and friend connections between users do not include timestamps. At times, collecting the data takes days or weeks or “forever” to complete, due to throttling or other technical limitation or the sheer size or the dynamic nature of source data.

Exploration. A visual analytics approach is key to enable users with varied technical skills to collaboratively explore and make sense of a phenomenon. Being able to follow the visual analytics approach requires process flexibility. That is, all the stakeholders of the analysis process should be able to conduct any of the individual steps by themselves even though development of the overall process requires technical development skills.

Transparency. Developers with technical skills may select to manage the network analysis data, in its different phases, with a database. To accomplish transparency and flexibility in the process, other members of the investigative team may, however, need less technical means to access the data. The use of intermediary results is key in facilitating the transparency and

flexibility of the process. Intermediary results refer to data in between the individual steps of the analysis. These data should be available as files in widely used formats, such as CSV and GEXF. In addition to the enhanced transparency, these intermediary results allow for speeding up the analysis process by using cached versions of source data and intermediary results when they have not changed.

Loose coupling. At best, data-processing pipelines can be built with a range of tools and components that have been built with different technologies. This kind of flexibility allows the introduction and use of new expressive tools from individual software components to full-featured applications as they become available to the investigative team. Many of them introduce new opportunities for advancing the analysis process but generally it is not possible to integrate these tools to a data-processing framework in program code (API) level.

Reproducibility. In the data-driven visual network analytics approach, reproducibility is first and foremost a technical quality of the process: the investigative team should be able to repeat the study or one or more of the analysis process and reproduce the results. Reasons for the need to rerun the process include, among others, updates on the source data, development steps of the analysis process, and the introduction of completely new processing steps and tools that insist on the use of a particular data format or extending the existing data. Moreover, dynamic sensemaking for complex phenomena mandates being able to refresh the data and derive new results with updated data. Reproducibility at this technical level also allows the investigative team to release the process, data and results to other researchers interested in the phenomena under investigation.

Enabling manual steps. While reproducibility is important, at the same time it is important to realize that automating some of the steps may not be feasible when an analysis is conducted the first time or requires intensive tailoring. Therefore, the process should support implementing any of the process steps manually. The use of file-based intermediary results is a practical approach in enabling manual analysis steps.

Automation. Allowing the development of automatically updating dashboards as needed gives the investigative team the opportunity to continue observing particular phenomena over time. It is expected that production-ready analysis processes for dashboards will operate without supervision; however, in the context of exploratory research, some requirements may be relaxed.

Low entry barrier. Analysis of innovation ecosystems and other network-based investigations of complex phenomena require extensive domain knowledge, and hence insist on active participation from domain experts (often without extensive technical expertise) throughout the analysis process. This requirement further underlines the need for transparency of the analysis process and the individual analysis steps.

Interoperability. The investigative team should be able to use a number of existing analytics tools with high usability and rich interactivity such as Gephi, NodeXL, KNIME and Tableau for conducting the analysis. Moreover, provisioning the visualized networks and other outputs of the analysis should be possible through dashboard built with Web technologies such as D3.js, DC.js, GEXF.js and the like.

In terms of the General Development Cycle, these requirements can be used to describe the Definition of the problem that serves as the starting point of artifact development (cf. Vaishnavi and Kuechler 2007). These requirements form a design rationale for the Ostinato exploration–automation cycles of the process model for data-driven network analysis.

4.2 Process model

The Ostinato process model that is presented in this section is developed over multiple case studies with a design research approach. It is built on existing models and previous work, and it takes into account the process requirements presented in section 4.1. Each step is described. Figure 1.2 shows a diagram of the process model.

Phase 1: Data Collection and Refinement

1. Entity Index Creation
2. Web/API Crawling
3. Scraping
4. Data Aggregation

Phase 2: Network construction and visualization

5. Filtering in Entities
6. Node and Edge creation
7. Metrics Calculation
8. Node and Edge Filtering
9. Entity Index refinement
10. Layout Processing
11. Visual Properties Configuration
12. Visualization Provision
13. Sensemaking, Storytelling & Dashboard Design

Phase 1: Data Collection and Refinement

The general rules of data-driven analytics apply here: collecting and cleaning the data will in most cases consume most of the time and resources available for the investigation.

Entity Index Creation

In some cases, the source data can be collected in full; whereas, in other cases only data on entities that are relevant for the analysis need to be collected. In one use case, we were interested in the Twitter discussions taking place in relation to a conference, #mindtrek. We collected all the Tweets sent by conference participants before, during and after the event in order to create a network representing the social structure of the conversation. For this, we created an entity index including the Twitter handles of conference participants, as well as those mentioned in the discussion (Jussila et al. 2014).

In the context of innovation ecosystem studies, the entities for which we collected data were defined by boundary specification (Basole et al. 2012). For example, in investigating the

connections between companies taking part in Young Innovative Companies program[2] run by the Finnish Funding Agency for Innovation Tekes, the list of companies defined the starting point of the analysis (Huhtamäki et al. 2012).

Web/API Crawling

Collecting the data is the most heterogeneous step in the data-driven visual analytics process. Possible source data potentially includes everything digital, from proprietary offline documents and document collections to spreadsheets to Web APIs (Application Programming Interface) to Web sites that are designed primarily for human interaction.

Similarly, the functionality required to collect the source data can range from relatively simple reading of individual documents to functions similar to a fully featured Web crawler. Compared to crawling random websites, Web APIs are, by default, more straightforward for data collection as they are often designed to support reuse (Vinoski 2008). At best, source data is available as linked data (Bizer et al. 2009), i.e. data that has a clear structure with individual facts that can be interconnected with the help of unique identifiers. This is key in ensuring referential integrity.

At the end of the crawling phase, a set of web resources, or rather their representations in Hypertext Markup Language (HTML) or some other format, is made available in a local database or other storage, a proxy that significantly speeds up the subsequent processing steps.

Scraping

Once the raw source data is available locally, the next step is to filter, select and distill the utility data relevant to the analysis process. Scraping refers to the process of distilling data from documents that are published to the Web for humans to use. Scraping can be seen as a form of the Extract, Transform, Load (ETL) process that is often applied in the context of data warehousing or other business intelligence processes to collect data from different sources to be refined and normalized and finally loaded into a consistent database for later use (Petschulat 2010, Vassiliadis 2009).

When collecting data from Wikipedia on Finnish Young Innovative Companies (YIC), for example, we were particularly interested in the facts presented in the Infobox section[3] of the page. To collect this data, we took advantage of the HTML markup on the page to specify the semantics (meaning) of the different pieces of text.[4] Each of the facts is represented as a table row including two cells, the first of which includes the label specifying the type of the fact and the second includes the actual value. Moreover, the value is also represented as a link to a separate page, a fact that we included in the crawl.

---

[2] Funding for young innovative companies, http://www.tekes.fi/en/funding/companies/funding-for-young-innovative-growth-companies/

[3] Help:Infobox, http://en.wikipedia.org/wiki/Help:Infobox

[4] The Terms of Service for a Web page must also be considered. When using Wikipedia as a data source, for example, one has to take into account the Terms of Service that specifically deny crawling Wikipedia for large amount of files. Instead of crawling the live website, users of the data are advised to download a copy of Wikipedia's contents and set up a proxy for serving further processing.

Data Aggregation

Social media studies often take place within the boundaries of an individual social media service; and therefore, ways of accessing data and identifying individual entities can be straightforward when one source of data is used. The complex context of innovation ecosystem studies, however, led us to use several sets of data in parallel. This meant that in many, if not most, of the cases, linked data was not readily available; and therefore, links between individual sets of data had to be created through finding unique entity identifiers that allow referential integrity. In innovation ecosystem studies, the name of the company or another actor is sometimes the key data point that can be used to identify an entity; in other cases, more advanced entity recognition procedures can be applied.[5] This kind of data cleaning is sometimes referred to as data wrangling (Kandel et al. 2011). Applying the methods of entity recognition provides a potentially more general solution to creating unique identifiers for entities in the data.

Phase 2: Network Construction and Analysis

Once the data is available on a local proxy, the utility data has been extracted from the source documents and data from different sources has been aggregated into a consistent set of linked data, the construction of the network representation of the phenomena under investigation can begin.

Filtering in Entities

The network construction phase starts by selecting the entities that will be included in the network. The selection of nodes is guided by the boundary specification designed and defined by the investigative team. At least two approaches exist to implement the selection: starting from a list of entities and rule-based entity inclusion. To continue the Finnish YIC example, we started from the list of companies participating in the program. We scraped Wikipedia data on the connections between the YIC companies and key individuals running them. If data on the individuals was not available in a clean format, we followed the crawling pattern by including the individuals in the list of web resources to be crawled. We continued to complement the dataset with data from the Innovation Ecosystems Network Dataset (IEN Startups and Angels, IEN Executives and Growth) and other sources of data about investments, acquisitions and affiliations.

A key reason to separate the selection of entities from node and edge construction is to support the transparency, reproducibility and extensibility of the process. To create a shared understanding of the analytical results, it is absolutely vital that all the investigators taking part in a particular network study are able to understand the original raw data, in addition to any constructed variables, and the various analytics and metrics that represent the network; this means that investigation participants need access, including access to the raw data. In our experience, we found that answering specific questions raised by anyone interested in the study, drawing conclusions, generalizing the results, developing more specific and potentially

[5] When using names as identifiers, one can apply fuzzy string matching and semi-automated tools such as OpenRefine (http://openrefine.org/) or DataWrangler (http://vis.stanford.edu/wrangler/) to assist in the aggregation process.

more interesting questions all depend on transparency of the data available and used for the analysis.

Node and Edge Creation

A key part of the data-driven network analysis process is, of course, the actual creation of the network. Network creation boils down to the creation of nodes representing the actors and the creation of edges representing the connections between the actors. Several options are available, however, when specifying details of the network creation process. First, the network can be either one-mode or two-mode. In one-mode networks all the nodes are of same type: startup companies, for example. Connections between the nodes are formed through relationships: investments, affiliations to individuals, acquisitions and transactions. In two-mode networks, there are two types of nodes, for example, startup companies and individuals related to them. Hypergraphs and bipartite graphs are examples of means to visualize two-mode networks (Freeman 2009, Jesus et al. 2009).

Further, the connections between network nodes can be either valued or dichotomous. With valued connections, the strength of a connection can be expressed. In either case, the connections may be undirected or directed. Finally, the temporal dimension can be included in networks if the data used to create the connections is time-stamped. With temporal data, insights about the evolution of the network can be gained.

Metrics Calculation

Network metrics enable quantifying a variety of structural properties, both in network and node level. These range from simple metrics such as node degree (indegree, outdegree) and betweenness to hub and authority values with HITS and other more sophisticated measures. Whereas in principle, every metric can be calculated for all of the networks and their nodes, in practice this is not feasible due to reasons of efficiency. Moreover, new metrics are being developed continually, and the investigative team is likely to find – or develop – new metrics that fulfill specific investigative purposes. From an implementation viewpoint, it is unlikely to find one tool that supports all the metrics the team wishes to use. Therefore, a combination of tools may be required to calculate the metrics.

As part of this step, network metrics for the network representation should be archived for later usage. For transparency, a list of exported network nodes and edges should include the various metrics used. In practice, node and network metrics must be recalculated after each change in the network structure; however, reference to previous calculations is often needed.

Nodes and Edge Filtering

A key limitation in visual network analysis is the amount of space available, both on screen and particularly on paper, to present the visualization. Depending on the level of detail required in the analysis, hundreds or thousands of nodes can be presented in one visualization view. For networks of tens of thousands of nodes and more, only more general structures and patterns can be observed from visualization. Two means exist to address this limitation: the best option is to allow the visualization users to filter in and out nodes and edges. If the end-user tools used to present the visualizations do not allow filtering, it can be done as one part of the automated process. Often, reducing the size of the visualized network is accomplished with a combination

of filtering out edges that have the least amount of weight as well as filtering out nodes that: (1) are left without edges; (2) have a value of the degree or some other a network analysis metric under a specified threshold; or (3) are (not) of particular type (even though this can already be taken into account when filtering in the entities used to construct the network in the first place).

-Entity Index Refinement

At this stage, the network is constructed and the required metrics are calculated for each of the nodes. Depending on the boundary specification applied in a particular investigation, the network is either ready to be visualized or, alternatively, additional data can be collected to complement the network. Revisiting the Finnish Young Innovative Companies case, the boundary specification was designed to include all the individuals involved in one or more of the companies in YIC program as well as all the other companies the individuals are or have been affiliated with. Moreover, the data included all the investors that had invested into any of the companies as well as all the companies that had acquired any of the YIC companies.

Layout Processing

The principle of processing network layout is simple. Nodes are given a position in two-dimensional space in a way that network structure is revealed in an intuitive way. Despite the simplicity, novel layout algorithms have continued to be developed over several decades. In our research cases, various stakeholders found a specific implementation of force driven layout, Force Atlas, to be particularly suitable for laying out networks representing innovation ecosystems at different levels. Force Atlas is implemented in Gephi and can be used as a batch process with the help of Gephi Toolkit[6]. In practice, the parameters of the layout algorithm must be adjusted manually for a particular kind of a network before fully automating layout processing. Alternatively, the layout can be processed with the UI version of Gephi and the resulting network, including the XY-coordinates for each node, can be exported, e.g. in GEXF.

Storing the network layout data is particularly important for improving the efficiency of the layout process, as well as for reducing investigators' cognitive load and promoting transparency. In particular, it is important that after the data is refreshed, the investigators are able to find the pre-existing nodes in an area of the network where the nodes were previously located. This stability can be achieved by inserting the existing positions into the network data before re-running the force driven layout algorithm. In most cases, investigators will find the pre-existing nodes close to the initial area of the network.

Future work is needed to determine how features such as layout algorithms, e.g., those implemented into NodeXL, could be used as a component of data-driven visual network analysis pipelines.

-Visual Properties Configuration

In networks, there are limited selection possibilities when defining the visual appearance of nodes and edges. Nodes have size, color and perhaps a border and shape as elected visual features. Edges have color and width. Allowing the user to select and change the visual properties according to node metrics and other node properties is perhaps the easiest way to

[6] Gephi Toolkit, http://gephi.github.io/toolkit/

allow end user interactivity in network analysis. Depending on the tools used by the investigators to conduct the analysis, the visual properties of nodes and edges can continue to be tweaked as part of the interactive analysis process.

Visualization Provision

At this stage, a network has all the required information available and therefore can be visualized. The means to finalize this step depend greatly on the tools that have been selected for use by the investigative team. In most cases, however, the created network is serialized into a file following a selected vocabulary or format for representing a network. These vocabularies and formats range from different CSV based applications to XML-based languages designed for representing networks.

A minimum approach to provision the network visualizations is to export network data in GEXF or other suitable format and place the resulting file into a folder from where a library such as Gexf.js can access it. More generally, viewer composition scenarios can include the following:

Scenario 1. Network viewer component with fixed functionality, i.e. following a fully descriptive approach. Visual properties such as node size and color need to be defined into the data during its processing. Gexf.js is an example of such a component that we have found useful in adding value to a fully static PDF-based approach in disseminating network visualizations.

Scenario 2. Implementing a dashboard with Web technologies, more specifically frameworks such as Highcharts, D3.js, Crossfilter.js, DC.js and others. In this case, tailored interactive features for data exploration can be provided to the user, adding options for representing network data.

Scenario 3. Using full-feature explorative analytics tools such as Gephi, NodeXL and Tableau, which can be used to further process the data and to connect source data to visual properties of the visualization. The key here is to produce visualizations rich-enough in data that the analyst can fully utilize the critical properties of the chosen analytics tool for investigation and exploration. In Gephi, for example, it is useful to include attribute data for nodes to assist network filtering in a way the investigator desires to do.

Sensemaking, Storytelling and Dashboard Design

While information visualization includes data transformation, representation, and interaction, it is ultimately about harnessing human visual perception capabilities to help identify trends, patterns, and outliers. Sensemaking has its roots in cognitive psychology and many different models have been developed. Sensemaking procedures are cyclic and interactive, involving both discovery and creation (North 2006). During the data collection and refinement phase, an individual searches for representations. In the network generation phase these representations are instantiated, and based in these insights the representation may be shifted, to begin the process again. Sensemaking is closely linked to the insight objectives (Konno et al. 2014), and the Ostinato cycle of exploration–automation is key in achieving actionable insights that network orchestrators can utilize.

When sensemaking requirements are satisfied for investigators and users, steps of the Ostinato process can be formalized with automated procedures for iteration over time. Key actors,

relationships and events of the network can be incorporated into dashboards that will track changes in critical assumptions and into stories that will share vision for actionable change.

5. Discussion

This second volume of the book presents new tools, applications, services, and algorithms needed to investigate how social media content is created, curated and disseminated and how the authority and trust of social media content creators and the impact of the content can be estimated. The Ostinato Model contributes to this call in two levels. First, it can be applied to support the data-driven investigations of innovation ecosystem structure and dynamics. Moreover, in the context of our investigations, social media serves first and foremost as a source of data that is fed into the investigations of innovation ecosystems and the structure between their actors. Therefore, second, for validity and reliability of these investigations, it is key to be able to increase the transparency of the processes behind these data originating from social media.

The Ostinato Model contributes to the data-driven network investigations of social media, innovation ecosystems and other network-driven phenomena in three different ways. First, the network approach has great strength in supporting the explorative studies of the patterns in between actors creating, curating and disseminating social media content. Second, referring specifically to the first phase of the Ostinato Model, data-driven approach allows tracking down processes over the boundaries of individual social media platforms and services. Third, user-centricity of the data-driven process adds to the transparency of the process itself, therefore providing means to triangulate different phases of data refinement and transformation and allowing different stakeholders of investigations to take as proactive role as they wish in moving forward a particular investigative process.

Due to the continued and rising interest in social media analytics and general big data analysis, new tools are continually introduced to support investigative work. Despite the tool development, a combination of tools is likely to continue to provide more flexibility in accessing and aggregating data and in processing and analyzing it. Finding a balance between user interface-operated low barrier tools and expressive computational strategies that require technical knowledge is key in making the investigative process as productive as possible while maintaining transparency and process flexibility.

This Ostinato Model for user-centric, process-automated, data-driven visual network analytics meets many of the requirements outlined earlier in this paper for the exploration–automation cycle recommended for developing shared understanding.

Setting up persistent data-collecting routines requires, in general, a programmatic implementation and must be designed and implemented case by case. To maintain the transparency of the process, it is important that the investigators are able to access both the raw data as well as to track down the various steps used to derive the data that is eventually used for the analysis and visualizations.

Allowing exploration boils down to the selection of the end user tools available for investigators to visualize and explore the data. If a rather static tool such as Gexf.js, for example, is used, the

user is limited to browsing and searching the data. If importing the data into an exploration platform such as Gephi or NodeXL is permitted, it is possible to provide the user with node and edge data, enabling them to continue their explorations with more technical independence. The availability of particularly expressive visual analytics tools, such as Tableau, adds to investigation options of analyzing network data, either as a network or using node and edge level data to provide new inspirations for other kinds of data analyses.

Using files rather than databases for representing intermediary results supports both loose coupling and transparency of the process. It also allows for implementing some of the steps manually, if seen feasible, and the flexibility of the process in general is increased.

Reproducibility is both a technical and a policy requirement. For an investigative team revisiting or extending an existing case, the availability of runnable code, source data and intermediary results provides a fruitful starting point. Moreover, results of reproducible studies can be published in a way that both data and runnable code are available, allowing a solid foundation for others to add their contributions as well. A reasonable proposition is that such a piece of knowledge draws attention from other researches and therefore has true potential for impact.

Automation is a key requirement for reproducibility, as well as for creating a dashboard that continues to update visualizations of the phenomena under investigation, sometimes in close to real time.[7]

Low entry barrier is enabled through making intermediary results available to all the members of the investigative team. As the process is repeatable and its individual steps are automated, new projections of the data can be implemented in an iterative and incremental manner. Implementing completely new steps of analysis becomes possible even without technical skills. Automating the steps, however, requires developers' attention. The Ostinato process model requires a multidisciplinary data science team or the somewhat mystical multi-skilled data scientist, (cf. Davenport 2014) to conduct the investigation.

Interoperability can be built into a computational approach. This requires that the technical architecture is flexible enough to permit different software components and tools – that may be implemented with different technologies – to be introduced into the process. When an analysis pipeline is built completely from scratch, it is recognizably important to minimize the number of technologies used. However, moving fast and in an agile manner is an objective we claim can be achieved when existing tools can be integrated to implement the individual steps of the analysis process and to provide the visualizations to investigators and other end users.

An implementation of the Ostinato user-centric, process-automated model for data-driven visual network analytics can serve as the core engine of an investigation. It can also be used to develop a pre-processing pipeline that collects and refines the data, creates a network representation and serializes the outputs to be analyzed and processed with expressive tools that, standing alone, allow the full visual analytics cycle for users.

[7] Using a full stack programming language such as Python gives the developers more opportunities to turn the scripts developed for analysis into processes that run in the cloud, intermittently collecting and preprocessing the data and feeding results into dashboards implemented in Web technologies.

6. Summary

In this chapter, we have presented the Ostinato Model of the exploration – automation cycle user-centric for data-driven visual network analytics. This model has two main phases, data collection and network analysis; they iterate through a cycle of exploration and automation. The Data Collection and Refinement step is divided into Entity Index Creation, Web/API Crawling, Scraping, and Data Aggregation. The Network Creation and Analysis step is composed of Filtering in Entities, Node and Edge Creation, Metrics Calculation, Node and Edge Filtering, Entity Index Refinement, Layout Processing, and Visual Properties Configuration. As a final step, the visualizations are provisioned to investigators and other end users with interactive exploration tools and discussion, and their feedback activates an iteration of the process. This Ostinato process model allows both an exploratory approach during the early phases of the investigation as well as the automation of the data collection and analysis process. The iteration cycle is especially beneficial in working with multi-source datasets, complex phenomena, changing externalities that may impact assumptions for decisions, and establishing a dashboard for continued observation of the phenomena over time, perhaps in real time.

A key challenge of this approach concerns the number of options for investigators and other end users to interact with the data in real-time while conducting the analysis, particularly the non-technical investigators on a multi-disciplinary team. The design research approach favors an iterative approach for both data-driven explorations and evidence-based decision making. However, investigators with limited programming skills or related technical know-how are limited in their participation, even though they may possess vital domain intelligence. Through access to data, documentation of changes in the analytical approach, flexible means to produce network representations in various formats, and exposition of intermediary results, barriers to participation can be lowered. The cycle of exploratory visual analytics, confirmation of data selection rules and analytical results made accessible through high interactivity visual analytics, allows the investigative team to confirm assumptions and investigative procedures, identify aspects of the analysis that can be automated and establish a transparent, replicable process.

The Ostinato process model has several implications for investigative teams taking the data-driven visual network analytics approach.

First, facilitation and documentation of the investigative process are required. Low barrier for entry in exploration and analysis poses risks that increase without transparency. Put another way, with added transparency and through intermediate results and easy access, the risk of false conclusions is lowered. Co-ordinated discussion on raw data and its journey to the finalized visualizations and other results is imperative; documentation of assumptions and rationale for changing data selection or analytical procedures enables transparency. Facilitation also helps in creating literacy of the processes and its outputs within the investigative team. Having the intermediate results available, all the members of the investigative team are able to maintain more of the control of the process and continue to introduce new, novel ways of analyzing the data as their skills and methodological know-how allows.

Second, the cycle of exploration – automation introduces new requirements for governance. Intermediary results require transparent authorship in their provenance. The transparent authorship of new datasets, constructed variables and analytical iterations must be ensured.

Third, starting from exploration and moving toward automation is straightforward with the help the process model. The investigative team is able to move fast in the beginning of the process while, at the same time, maintaining control over the process as its complexity increases. With appropriate technology selection, the process can eventually be relegated to the background to collect, process, analyze and visualize data in an automated manner to support a longitudinal study of a particular phenomena. And, more importantly, a mature procedure – or one or more of its components – can be reused to investigate other phenomena of interest.

Fourth, increased reproducibility is an asset for future studies but requires explicit governance. Technical reproducibility of the process allows revisiting analytical results of a case even after a long time period. Refreshing (collecting new) data or, alternatively, adding new dimensions into existing data is straightforward when the process or its individual parts can be run computationally. Curational rules must be developed, and access to code and data has to be designed at both the technical and policy levels. Governance of the data from raw to intermediate results to outputs as well as the components and software process must be articulated.

Within the constraints imposed by the level of abstraction in this article, this Ostinato process model provides blueprints for designing analytical processes with technologies ranging from Python to R to Javascript. At best, the process is able to support the inclusion of several different technologies, as implemented e.g. by the Wille Visualisation System (Nykänen et al. 2008).

Future work includes, first, the refinement of this model on basis of the feedback collected from researchers and practitioners working with the exploration – automation cycle of data-driven visual network analytics and applying the model and, second, the implementation of a software framework – perhaps similar to Grunt (http://gruntjs.com/), a popular Javascript-based task runner – to support the development of processes of data-driven visual network analytics on very large datasets.

As an ecosystem of tools and components develops and requirements for interoperability are articulated, we see the possibility of developing a community of people moving the field forward. They will need a package management framework, system components and a supportive community.

The Kredible.net initiative is an important step toward establishing a community like this.

Acknowledgements

The research reported in this chapter was funded through resources provided by Tekes – the Finnish Funding Agency for Innovation and mediaX at Stanford University.

References

Barabási, A.-L. & Bonabeau, E., 2003. Scale-Free Networks. *Scientific American*, 288(5), pp.50–59.

Basole, R.C. et al., 2012. Understanding Mobile Ecosystem Dynamics: A Data-Driven Approach. In *Proceedings of the 2012 International Conference on Mobile Business (ICMB 2012)*. Delft, Netherlands, pp. 17–28. Available at: http://aisel.aisnet.org/icmb2012/15/ [Accessed February 7, 2013].

Bizer, C., Heath, T. & Berners-Lee, T., 2009. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), pp.1–22.

Card, S.K., Mackinlay, J.D. & Shneiderman, B., 1999. *Readings in information visualization: using vision to think*, San Francisco, Calif.: Morgan Kaufmann Publishers. Available at: http://www.amazon.com/Readings-Information-Visualization-Interactive-Technologies/dp/1558605339.

Freeman, L.C., 2009. Methods of Social Network Visualization. In Berlin: Springer.

Freeman, L.C., 2000. Visualizing Social Networks. *Journal of Social Structure*, 1(1), p.[np]. Available at: http://www.cmu.edu/joss/content/articles/volume1/Freeman.html.

Ghosh, S., 2013. *Python Tools for Reproducible Research in Brain Imaging*, Available at: https://speakerdeck.com/satra/pydata-2013-python-tools-for-reproducible-research-in-brain-imaging.

Giuliani, E. & Bell, M., 2008. *Industrial clusters and the evolution of their knowledge networks: revisiting a Chilean case*, Falmer, Brighton, UK. Available at: http://www.sussex.ac.uk/spru/documents/sewp171.

Granovetter, M., 1973. The Strength of Weak Ties. *American journal of sociology*, 78(6), pp.1360–1380. Available at: http://www.jstor.org/discover/10.2307/2776392?uid=3737976&uid=2&uid=4&sid=21104852601921.

Hansen, D. et al., 2009. *Do you know the way to SNA?: A process model for analyzing and visualizing social media data*, College Park, Maryland. Available at: http://www.cs.umd.edu/localphp/hcil/tech-reports-search.php?number=2009-17 [Accessed July 3, 2014].

Hansen, D., Shneiderman, B. & Smith, M.A., 2011. *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*, Burlington, MA, USA: Morgan Kaufmann. Available at: http://www.amazon.com/dp/0123822297.

Heer, J. & Shneiderman, B., 2012. Interactive Dynamics for Visual Analysis. *Communications of the ACM*, 55(4), pp.45–54. Available at: http://dl.acm.org/citation.cfm?id=2133821 [Accessed January 31, 2013].

Hevner, A.R. et al., 2004. Design Science in Information Systems Research. *MIS Quarterly*, 28(1), pp.75–105.

Huhtamäki, J. et al., 2010. Context-Driven Social Network Visualisation: Case Wiki Co-Creation. In D. Karabeg & J. Park, eds. *Proceedings of the Second International Workshop on Knowledge Federation: Self-Organizing Collective Mind, Dubrovnik, Croatia, October 3-6, 2010.* Dubrovnik, Croatia: CEUR-WS.org. Available at: http://urn.fi/URN:NBN:fi:tty-201201161008.

Huhtamäki, J. et al., 2012. Networks of Growth: Case Young Innovative Companies in Finland. In *Proceedings of the 7th European Conference on Innovation and Entrepreneurship (ECIE), September 20-21, 2012, Santarém, Portugal*. Santerém, Portugal.

Huhtamäki, J. et al., 2013. Process for Measuring and Visualizing an Open Innovation Platform: Case Demola. In *17th International Academic MindTrek Conference 2013: "Making Sense of Converging Media", October 1-3, Tampere, Finland*. ACM. Available at: http://urn.fi/URN:NBN:fi:tty-201312201533.

Hwang, V.W. & Horowitt, G., 2012. *The Rainforest: The Secret to Building the Next Silicon Valley* 1.02 editi., Los Altos Hills, Calif.: Regenwald. Available at: http://www.amazon.com/The-Rainforest-Secret-Building-Silicon/dp/0615586724.

Indarto, E., 2013. Data Mining. Available at: http://recommender-systems.readthedocs.org/en/latest/datamining.html [Accessed December 13, 2014].

Intel, 2013. *Extract, Transform, and Load Big Data with Apache Hadoop*, Available at: https://software.intel.com/en-us/articles/extract-transform-and-load-big-data-with-apache-hadoop.

Jesus, R., Schwartz, M. & Lehmann, S., 2009. Bipartite networks of Wikipedia's articles and authors: a meso-level approach. In Orlando, Florida: ACM, p. Article 5, 10 pages. Available at: http://portal.acm.org/citation.cfm?id=1641309.1641318 [Accessed September 2, 2010].

Jussila, J. et al., 2014. Visual network analysis of Twitter data for co-organizing conferences: case CMAD 2013. In *Proceedings of the 47th Annual Hawaii International Conference on System Sciences, January 6-9, 2014*. Computer Society Press, pp. 1474–1483. Available at: http://urn.fi/URN:NBN:fi:tty-201401221053.

Kandel, S. et al., 2011. Research Directions in Data Wrangling: Visuatizations and Transformations for Usable and Credible Data. *Information Visualization*, 10(4), pp.271–288. Available at: http://dx.doi.org/10.1177/1473871611415994 [Accessed October 8, 2014].

Keim, D., Kohlhammer, J. & Ellis, G. eds., 2010. *Mastering the Information Age - Solving Problems with Visual Analytics*, Eurographics Association. Available at: http://www.vismaster.eu/book/.

Konno, N., Nonaka, I. & Ogilvy, J., 2014. Scenario Planning: The Basics. *World Futures*, 70(1), pp.28–43. Available at: http://www.tandfonline.com/doi/abs/10.1080/02604027.2014.875720 [Accessed December 14, 2014].

Leskovec, J., Backstrom, L. & Kleinberg, J., 2009. Meme-tracking and the dynamics of the news cycle. In KDD '09. New York, NY, USA: ACM, pp. 497–506. Available at: http://doi.acm.org/10.1145/1557019.1557077 [Accessed September 14, 2012].

Liu, Y.-Y., Slotine, J.-J. & Barabási, A.-L., 2011. Controllability of complex networks. *Nature*, 473(7346), pp.167–173. Available at: http://dx.doi.org/10.1038/nature10011 [Accessed June 15, 2011].

Mawer, D., 2000. Ballet and the Apotheosis of the Dance, In Mawer, D. (Ed) *The Cambridge Companion to Ravel*, Cambridge University Press: Cambridge, 157.

Moreno, J.L., 1953. *Who Shall Survive?: Foundations of Sociometry, Group Psychotherapy and Sociodrama*, Beacon, NY, USA: Beacon House Inc. Available at: http://www.asgpp.org/docs/WSS/WSS.html [Accessed August 23, 2010].

North, C., 2006. Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, 26(3), pp.6–9. Available at: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1626178&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D1626178.

Nykänen, O. et al., 2008. A Visualisation System for a Peer-to-Peer Information Space. In Tampere, Finland: Tampere University of Technology, pp. 76–85. Available at: http://matriisi.ee.tut.fi/hypermedia/events/opaals2008/articlelist.html#opaals2008-article14.

Paquin, R.L. & Howard-Grenville, J., 2013. Blind Dates and Arranged Marriages: Longitudinal Processes of Network Orchestration. *Organization Studies*, 34(11), pp.1623–1653. Available at: http://oss.sagepub.com/content/34/11/1623 [Accessed March 11, 2014].

Peffers, K. et al., 2007. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3), pp.45–77. Available at: http://dx.doi.org/10.2753/MIS0742-1222240302 [Accessed October 29, 2014].

Peng, R.D., 2009. Reproducible research and. *Biostatistics*, 10(3), pp.405–408. Available at: http://biostatistics.oxfordjournals.org/content/10/3/405 [Accessed December 14, 2014].

Petschulat, S., 2010. Other people's data. *Communications of the ACM*, 53(1), p.53. Available at: http://cacm.acm.org/magazines/2010/1/55742-other-peoples-data/fulltext [Accessed December 14, 2014].

Pirolli, P. & Card, S., 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*.

Ritala, P., Armila, L. & Blomqvist, K., 2009. Innovation orchestration capability — Defining the organizational and individual level determinants. *International Journal of Innovation Management*, 13(04), pp.569–591. Available at: http://www.worldscientific.com/doi/abs/10.1142/S136391960900242X [Accessed March 6, 2014].

Ritala, P. & Hallikas, J., 2011. Network position of a firm and the tendency to collaborate with competitors – a structural embeddedness perspective. *International Journal of Strategic Business Alliances*, 2(4), pp.307–328. Available at: http://dx.doi.org/10.1504/IJSBA.2011.044859 [Accessed December 19, 2014].

Ritala, P. & Huizingh, E., 2014. Business and network models for innovation: strategic logic and the role of network position. *International Journal of Technology Management*, 66(2), pp.109–119. Available at: http://dx.doi.org/10.1504/IJTM.2014.064608 [Accessed December 19, 2014].

Rubens, N. et al., 2011. Alumni network analysis. In Amman, Jordan: IEEE, pp. 606–611. Available at: http://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=5773200.

Russell, M.G. et al., 2015. Relational Capital for Shared Vision in Innovation Ecosystems. *Triple Helix: A Journal of University-Industry-Government Innovation and Entrepreneurship*, forthcoming.

Russell, M.G. et al., 2011. Transforming Innovation Ecosystems through Shared Vision and Network Orchestration. In *Proceedings of Triple Helix IX International Conference: "Silicon Valley: Global Model or Unique Anomaly?", July 2011, Stanford, California, USA*. Stanford, California, USA.

Schwaber, K. & Beedle, M., 2001. *Agile Software Development with Scrum*, New Jersey: Prentice Hall.

Shakarian, P., Eyre, S. & Paulo, D., 2013. *A Scalable Heuristic for Viral Marketing Under the Tipping Model*, Available at: http://arxiv.org/abs/1309.2963 [Accessed September 23, 2013].

Simon, H.A., 1969. *The Sciences of the Artificial*, Cambridge, MA: MIT Press.

Still, K. et al., 2014. Insights for orchestrating innovation ecosystems: the case of EIT ICT Labs and data-driven network visualisations. *International Journal of Technology Management*, 66(2/3), pp.243–265.

Still, K. et al., 2013. Networks of innovation relationships: multiscopic views on Finland. In *Proceedings of the XXIV ISPIM Conference – Innovating in Global Markets: Challenges for Sustainable Growth, 16-19 June 2013, Helsinki, Finland*. p. 15.

Telea, A.C., 2008. *Data visualization: principles and practice*, Wellesley, Mass.: A K Peters. Available at: http://www.amazon.com/Data-Visualization-Principles-Alexandru-Telea/dp/1568813066.

Thomas H. Davenport, 2014. *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*, Boston, Massachusetts, USA: Harvard Business Press Books.

Thomas, J.J. & Cook, K.A., 2006. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1), pp.10–13.

Vaishnavi, V.K., Kuechler, W. & Jr, W.K., 2007. *Design Science Research Methods and Patterns: Innovating Information and Communication Technology*, Boca Raton: Auerbach Publications.

Vassiliadis, P., 2009. A Survey of Extract–Transform–Load Technology. *International Journal of Data Warehousing and Mining*, 5(3), pp.1–27.

Vinoski, S., 2008. Serendipitous Reuse. *Internet Computing*, (January/February 2008), pp.84–87. Available at: http://steve.vinoski.net/pdf/IEEE-Serendipitous_Reuse.pdf [Accessed February 27, 2009].

Wasserman, S. & Faust, K., 1994. *Social Network Analysis: Methods and Applications* 1st ed., New York, NY, USA: Cambridge University Press. Available at: http://www.amazon.com/dp/0521387078.

Weng, L., Menczer, F. & Ahn, Y.-Y., 2013. Virality prediction and community structure in social networks. *Scientific reports*, 3, p.2522. Available at: http://www.nature.com/srep/2013/130828/srep02522/full/srep02522.html [Accessed March 20, 2014].

Wikipedia, 2014. Web crawler. *Wikipedia*. Available at: http://en.wikipedia.org/w/index.php?title=Web_crawler&oldid=635502147 [Accessed December 14, 2014].