Informatics

Events Analysis in Visual Analytics

PhD THESIS

submitted in partial fulfillment of the requirements for the degree of

Doctor of Technical Sciences

within the

Vienna PhD School of Informatics

by

Roger A. Leite, M.Sc.
Registration Number 01528045

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.-Prof. Ph.D. Silvia Miksch
Second advisor: FH-Prof. Priv.-Doz. Ph.D. Wolfgang Aigner

External reviewers:
Associate Professor Ph.D. Remco Chang, Tufts University, USA
Univ.-Prof. Ph.D. Tatiana von Landesberger, Universität zu Köln, Germany

Vienna, 7th May, 2021

Roger A. Leite	Silvia Miksch

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Declaration of Authorship

Roger A. Leite, M.Sc.

I hereby declare that I have written this Doctoral Thesis independently, that I have completely specified the utilized sources and resources and that I have definitely marked all parts of the work - including tables, maps and figures - which belong to other works or to the internet, literally or extracted, by referencing the source as borrowed.

Vienna, 7th May, 2021

Roger A. Leite

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Acknowledgements

Thanks...

- To my mother and father, the ones who born and shaped me. You were my biggest and best inspiration.
- To my affective families for sharing part of the life with me through friendships.
- To Silvia Miksch, with whom I feel blessed to have worked. Silvia taught me not only about highly technical domain knowledge but also about personal skills. From the beginning, Silvia believed in a strange guy from the other side of the planet and supported all the migration challenges that came to happen.
- To the entire CVAST group for being great colleagues who taught and inspired me in many different ways along my research journey. Special thanks to Markus Bögl and Theresia Gschwandtner.
- To all professors, co-authors, reviewers, and fellow scientists involved in my learning and development. Thanks for sharing time and knowledge with me.

I share this victory with you all.

Finally, the research described in this dissertation was supported and funded by the programs: the Austrian Science Fund (FWF), KnoVA – Knowledge–Assisted Visual Analytics (FWF, Project #P31419-N31); The research unit "Visual Analytics" Centre for Visual Analytics Science and Technology (CVAST, #193-07); PolyCube (FWF, Project #P28363-24); and Lower Austrian Fund for Research and Education (NFB, Project SC16-032).

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Kurzfassung

Die Datenproduktion ist in bestehenden Domänen in ständigem exponentiellem Wachstum. Das Verständnis von Big Data wird zu einem zentralen Wettbewerbsaspekt. In jüngster Zeit werden Eigentum und Nutzung von Daten angesichts ihrer Macht und möglicher Folgen für die Gesellschaft auf internationaler Ebene diskutiert. Der volle potenzielle Einfluss bleibt unbekannt. Das Bewusstsein für Daten kann Menschen von der persönlichen bis zur staatlichen Ebene nutzen. Erfolgreiche Datenanalysen unterstützen die Entscheidungsfindung in verschiedenen Skalierungen und Lebensbereichen. Unabhängig von der Domäne und Skala werden die meisten Daten unter multivariaten und zeitorientierten Aspekten gesammelt und behandelt.

Die Ereignisanalyse berücksichtigt Variablen, die ihr Verhalten über die Zeit verändern. Die Identifizierung von Datenmustern und Datenanomalien und die Schlussfolgerungen daraus unterstützen kritische Aufgaben in verschiedenen Domänen. Derzeitige Lösungen für die Ereignisanalyse verwenden hauptsächlich Data-Mining-Ansätze. Die Anwendung von Visual Analytics (VA)-Techniken kann jedoch den Prozess der Wissensentdeckung verbessern und die Genauigkeit der Erkennung und Vorhersage von Ereignissen erhöhen. Durch verschiedene Ansichten und Perspektiven auf die Daten und die interaktive Exploration ermöglichen VA-Techniken dem Benutzer*innn, sich mit den Daten vertraut zu machen, während diese erkundet werden. Durch die Kopplung von menschlichen visuellen Wahrnehmungsfähigkeiten und Domänenwissen bietet VA verbesserte kognitive Vorteile.

In dieser Arbeit präsentieren und untersuchen wir wie VA angewendet werden kann, um die wichtigsten Herausforderungen in der Ereignisanalyse zu bewältigen. Die Hauptbeiträge dieser Arbeit sind: (1) wir haben verschiedene VA-Ansätze in enger Zusammenarbeit mit Expert*innen aus verschiedenen Domänen entwickelt, um reale Datensätze zu analysiernen und Aufgaben der Ereignisanalyse aus ihrem bestehenden Workflow zu verbessern, (2) wir stellen den ersten VA-Ansatz vor, der auf einem Scoring-System für die Erkennung von Finanzbetrugsereignissen basiert, (3) wir bieten eine neue, mit Anleitungen ("guidance") angereicherte Komponente für die Generierung, Erkennung und Filterung von Netzwerkmustern an, die verschiedene Ebenen der Analysekomplexität unterstützt, (4) wir haben verschiedene Evaluierungen unserer Lösungen durchgeführt, die positive Ergebnisse gezeigt haben, und (5) wir skizzieren mögliche zukünftige Forschungsrichtungen und offene Herausforderungen des Feldes. Alle unsere Entdeckungen wurden durch die

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

kontinuierliche Zusammenarbeit mit verschiedenen Domänenexpert*innen während des Entwurfs, der Entwicklung und der Evaluierung der einzelnen Experimente gesammelt.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Abstract

Data production is in constant exponential growth in various domains. The understanding of big data becomes a central competitive aspect. Recently, ownership and usage of data have been discussed on an international scale considering its power and possible consequences for society. Its full potential impact remains unknown. Data awareness can benefit people from the personal to the governmental level. Successful data analysis support decision-making in different scaling and life aspects. Regardless of the domain and scale, most data is being collected and treated present multivariate and time-oriented aspects.

Event analysis takes into consideration variables that change behavior over time. Data pattern and data anomalous identification and the reasoning about it support critical tasks among various domains. Currently, event analysis solutions use mainly data mining approaches. However, applying Visual Analytics (VA) techniques may enhance the knowledge discovery process and increase the detection and prediction of events' accuracy. As displaying distinct data perspectives in multiple views and with interactive support, VA aspects allow users to get familiar with the data while exploring it. By coupling human visual perception skills and domain knowledge, VA presents improved cognitive advantages.

We propose to investigate how VA can be applied to tackle the main challenges in event analysis. The main contributions of this thesis are: (1) we developed distinct VA approaches in close collaboration with experts from different domains to support real-world datasets and improve event analysis tasks from their existing workflow, (2) we present the first VA approach based on a scoring system for financial fraud events detection, (3) we offer a new guidance-enriched component for network pattern generation, detection, and filtering that supports different levels of analysis complexity, (4) we conducted different evaluations of our solutions that presented positive results, and (5) we elaborate on possible future research directions and open challenges in the field. All of our discoveries have been collected through continuous collaboration with different domain experts during each experiments' design, development, and evaluation.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Contents

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Part I

Overview

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

CHAPTER 1

Introduction

1.1 Motivation and Problem Statement

Time is a unique dimension that is present in the majority of data science works. Many domains benefit from time-oriented data collection and analysis. Time-oriented data analysis is a complex task that can be divided into different aspects. The detection and analysis of events are main tasks of time-oriented data analysis among many domains.

Regardless of the domain, when involving temporal data, the identification and understanding of temporal events allow several advantages to problem management and decision making. Artificial Intelligence (AI) algorithms and fraud detection metrics are currently used to identify anomalous events in different fields [Faw+98; HLN09; Gau+13; Cha+07; Sak+10; ASS13]. However, real-world data can be considered a natural phenomenon that is continually changing to generate new action that does not fit into known patterns. In other words, real-world data is volatile, flexible, and uncertain. Thus, search automatically for known patterns is not enough. Cook and Thomas define Visual Analytics (VA) as the "science of analytical reasoning facilitated by interactive visual interfaces" [TC05](p.4). These methods have the potential to improve new pattern discovery. Since real-world data is continually being changed, new "automatic rules" need to be constantly updated. VA can aid experts with these tasks by binding human features such as good pattern recognition and visual outlier perception with computational power. Referring to the two domains presented in this thesis, examples of event detection and analysis tasks are:

Bank Transactions. Financial fraud detection (FFD), identifies financial frauds types (money laundry, straw person, internal), fraud detection algorithm fine-tuning, a decrease of false-positive and false-negative alarms, and management of fraudulent events.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Exchange Supply Chain. Exploration of seasonal flows between sectors and regions, understanding how different economic sectors relate, support of network exchange profile creation, and supply chain exploration.

In this Ph.D. thesis, we present three different publication works of Roger‘s first authorship, which consider event analysis in two other domains (Bank Transactions and Exchange Supply Chain). All results are presented, discussed, and linked to present different scenarios where the VA could support temporal event-related tasks. The main contributions are:

- The presentation of a novel VA approach based on a scoring system for fraud detection (EVA).
- The development of a novel VA approach with sophisticated interactive network visualization that presents integrates guidance aspects (NEVA).
- The presentation of guidance-enriched components for network profile generation, network pattern detection, and filtering (Hermes).
- All works presented complete evaluations focused on insights tracking from real-world domain experts.
- The elaboration on findings and future research.

1.2 Background

In this section, we contextualize the thesis. To do so, we present the background for (1) Data Analysis, Data Visualization and Visual Analytics, (2) events, and (3) events focusing on frauds.

1.2.1 Data Analysis, Data Visualization, and Visual Analytics

This section explores a brief history and the definition of three terms: Data Analysis, Data Visualization, and Visual Analytics.

Data Analysis. Tukey [Tuk77] introduces the field of exploratory data analysis. Instead of focusing on create hypotheses and test it against the data, the author also considered the benefits that the exploration freedom of the data could bring. The main focus was to look at the data (graphically or numerically) and explore it before generating hypotheses that, alongside, can be tested and further confirmed/reject. This contribution is significant and still very useful to data science. Nowadays, we have improved a lot the way that we can look and interact with data. However, Tukey‘s contributions are still very central.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Data Visualization. Cleveland [Cle93] argues that data visualization always existed in different forms and shapes through other domains of science and technology. Data Visualization is also called just “Visualization”. The textbooks about statistical time series analysis [CC08], the literature of graphical representations [RG13; Fun36; Fun37], and the literature of information visualization [TG83; Aig+11] all point to a graph that should represent a inclination plot of planetary orbits over time [Fun36] as the first time-series graph found in literature. It dates to the 10th century (see Figure 1.1).

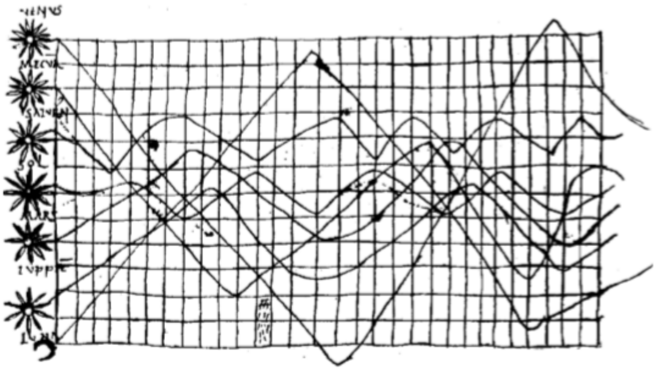

Figure 1.1: Possibly the first time series graph found in literature (10th or 11th century), author unknown [Fun36]. This graph shows a function of time to planetary orbits. Figure Source: Funkhouser (1936) [Fun36], courtesy of the ©University of Chicago Press.

The next register after this graph was 800 years later [Edw83]. The works of William Playfair and Johann Heinrich Lambert, in the late of 1700s, conceived them the honor of being considered the inventors of modern graphical design. Besides visualizing the data, they could also target future events by discovering and analyzing economic trends [Fun37] (Playfair). Moreover, the detection of a periodic variation and its period’s determination was also part of Lambert experiments [Til75]. “It was an early lesson that we gain a lot by plotting the points.” [Tuk77](p.665).

Visual Analytics. The research field of visual analytics is guided by the demand to cope with and make sense of the massive amount of complex collected data increased due to technological advancement. By using visual analytics, researchers could contribute to data analysis problems in many domains. The term “visual analytics” and its scope (see Figure 1.2) was first defined in 2005 by Thomas and Cook [TC05](p.4): “The science of analytical reasoning facilitated by interactive human-machine interfaces”.

This first book was motivated by homeland security against terrorist threats. Furthermore, the scope and challenges of visual analytics were defined by Keim et al. [Kei+08], where the first significant contributions were elaborated. This work adapted and extended the famous information-seeking mantra “overview first, zoom/filter, details on demand” [Shn96] for visually exploring data: “analyze first - show the important - zoom, filter and analyze further - details on demand” [Kei+08]. This adaptation was necessary

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

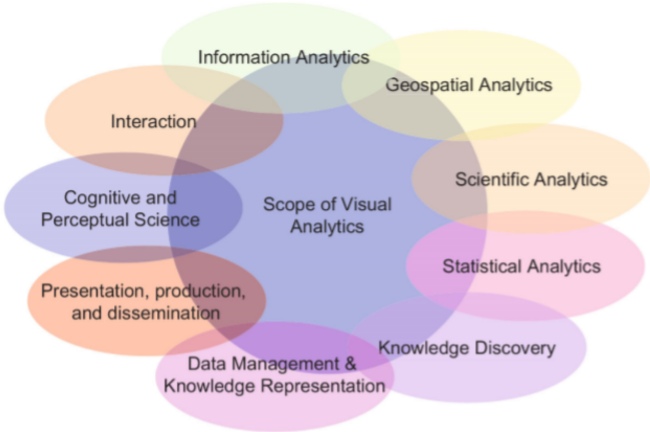

Figure 1.2: The scope of visual analytics defined by Keim et al. [Kei+08]. Figure Source: Keim et al. (2008) [Kei+08], courtesy of ©Springer 2008.

due to the visual aggregation information loss, a natural phenomenon during transcribing raw data to visual representations (especially for massive datasets). The authors argue for having a first analysis looking for exciting data aspects, show what seems to be necessary, allow interactions (zoom, filtering, etc.), and narrow the analysis until we get to the details on demand.

1.2.2 Events

Events are semantic occurrences in a particular time, place, and context. Events come in many forms and shapes according to the domain. It can be cyclical or sporadic, predictable or unexpected, constant, or mutable. Events may also work as time references: "something happens after/before a certain event". The detection of events is an essential task in many domains, such as detecting interesting changes in stock markets, detecting problems in health parameters, or detecting financial fraud. Analyzing these events in a temporal context fosters further insights such as frequency, trends, and changes. Once an event is identified, it can be classified, which is usually done in a domain-specific way.

Temporal event detection is a broad subject. It is relevant in diverse fields, such as biology, security, finances, sales, social networks, and disease monitoring. One recent example is the survey provided by Atefeh, et al. [AK13] that shows techniques for event detection from Twitter streams. The authors address the problem of analyzing Twitter content, and they classify the existent techniques by event type, detection task, and detection method. Guided by text stream visualizations, another example of an event detection survey is presented by Šilic, et al. [ŠB10]. In this article, a new method comparison by data type, text representation, and temporal drawing approach is presented.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

1.2.3 Events: focused on frauds

While event detection is aimed at identifying any type of event (not necessarily anomalous events), outlier detection focuses on patterns and samples that do not conform to expected behaviour, i.e. anomalies or outliers [CBK09]. In the case of fraud detection, for example, it is interesting to observe and classify new data change events as harmful or not. For instance, the purchase of a car might cause the transaction of a high amount of money that is unusual for the respective bank account, and thus, it might be classified as an outlier transaction. Such transactions of high amounts of money require special attention before being executed in order to avoid fraudulent schemes, such as hackers trying to use someones credit card for their own benefit. However, not every fraudulent event can be classified as an outlier, sometimes attacks are hidden in known patterns in order to avoid detection by simple rule-based approaches. The well renowned Oxford Dictionary defines fraud as "wrongful or criminal deception intended to result in financial or personal gain" [1].

Besides its challenging nature, visual fraud event detection has also a strong social and financial importance. For instance, fraudulent schemes such as 'money laundering', or 'straw person' should be detected and fought as fast as possible by financial systems. Governments, banks, and other financial institutions that provide credit and money transaction services are always interested in improving operation monitoring and fraud detection. Software environments handling sensitive data such as financial operation management systems, systems for insurance evaluation, or companies' internal control systems, need to be in constant evaluation to detect ever-changing fraudulent attempts, to provide risk management, and, thus, to avoid catastrophic consequences.

Most existing approaches in the field of fraud detection tries to represent sequences of suspicious events identified by AI and metric techniques [LF10]. In these cases, visualization techniques are not being fully exploited, it works just as a simple output filter. In addition, these approaches are only able to identify events that have already happened, being insufficient to investigate "on going"/"live" data. One of the main challenges of using AI techniques to detect frauds in real-world data is the constant adaptions that the system may require due to the ever-changing nature of the fraudulent strategies. This problem feature leads such "automatic-only" approaches to increase their false-negative alarms.

Concerning VA for event detection: focusing on frauds, an entire state of the art research was done by Roger Leite et al. [Lei+18b]. In this work, the authors collect fraudulent events detection solutions through five main domains: the stock market, banks, telecommunication companies, insurance companies, and internal frauds. The work classifies approaches by application domains, visualization techniques, and interaction designs, and analytical methods. The authors describe and argue about each system according to the classification schemes and discuss the main challenges for future research topics.

[1] http://www.oxforddictionaries.com/definition/english/fraud (accessed November 4, 2020)

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

1.2.4 Events: focused on exchange supply chain

Looking for anomalies or outliers [CBK09] is not enough when it comes to analyzing events in a supply chain network. The domain presents a wide range of events interpretation. For example, the increase in a certain sector of a given period within a region is considered an economic event. However, the opening of a new company, on its own, may also be considered an event possibility.

The supply exchange between different economic sectors presents all range of events interpretation as the involved domains allow. For example, a water crisis may show an industrial "domino effect" since water is an essential resource for factory production. With Hermes, we would be able to analyze the "water crisis" as a single event by overlooking the different regions and following individual companies' effects. As the "water crisis" example, many others may rise given the real-world non-deterministic feature. Therefore, the event classification is unlimited.

The exploration of an exchange supply chain dataset is a playground for event interpretation and a powerful investment analysis tool. Being aware of network features may give different users a competitive advantage (e.g., urban planners, investors, individual sector strategists). As an example, Tekušová et al [TK08] supports the identification of shareholders movement. The information of the entrance or exit of main players may be considered an important event and guide investors' decisions as well as knowledge of the increase/decrease of exchange by sector or region (feature presented by Hermes).

In Hermes [Lei+20b], we present different economic scales that also increase the interpretation and possibilities of the events. Hermes empowers users to analyze the exchange flow in two levels: (1) the user can explore sector and regional features of the exchange flow, and (2) the user can analyze the data at a company level.

1.3 Methodology

In the following section, we present the research methodology consisting of the design triangle, the nested model, and the derived research workflow.

1.3.1 Data-User-Tasks Approach.

Following the design triangle [MA14b] (see Figure 1.3), to generate interactive VA methods, we first tackle the three main aspects: data, users, and tasks. Each of these main features implies in a different quality criterion, as we can see in Figure 1.3.

Next, we present an example of how the discrimination between data, user, and tasks was used to support EVA [Lei+17] and NEVA [A L+20], two of our main works within a bank transaction domain.

Data: Financial transaction events constitute multivariate and time-oriented data, which may include details regarding to the accounts, types of transactions, amount of money, etc.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

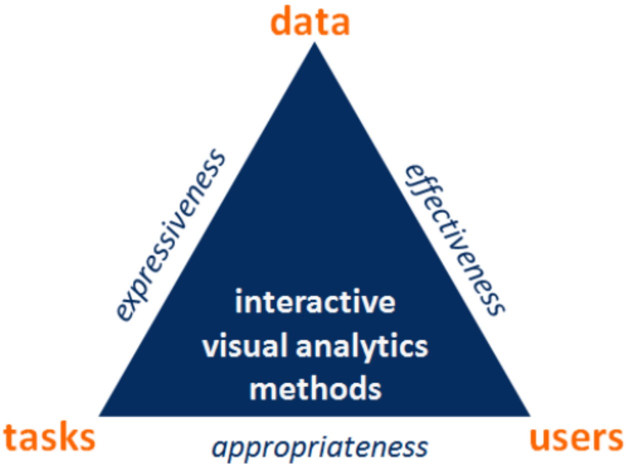

Figure 1.3: Data-User-Tasks Triangle to Visual Analytics. Figure source: [MA14a], ©2014 Elsevier. Used with Permission

Users: We may consider three different types of users. (1)Fraud transaction managers are non-expert users interested in looking into fraudulent transactions alarms, evaluate, and understand why one or more trades were stopped. (2) Investigators are interested in maintaining the system by identifying reasonable thresholds, fine-tuning scores, and pattern detection. (3) Local investigators are the ones between the fraud transaction manager and investigator who test "what-if" scenarios in the data and monitor the account's history.

Tasks: We identified the following tasks: fraud detection, fraud classification, fraud detection algorithm fine-tuning, credibility analysis, and customer behavior monitoring.

Next, we present another example of the data, user, and task design approach. This discrimination was used in Hermes [Lei+20b]. Hermes is a work within the exchange supply chain domain.

Data: Hermes extracts relevant information about the general economic features and behavior (in terms of investments) of specific regions and helps reconstruct the supply chain for expert analysis starting from two categories of data. (1) The first category describes aggregate information, e.g., on a national level, that reports the monetary in & out (IO) flows between different industry sectors within a fiscal year. This is known as "IO Tables" [Tim+15]. A transaction in such a table describes, for example, the amount of investments (in millions of Euros) that went from the "Water" sector to the "Agriculture" sector or the "Manufacturing" sector. Sectors are classified and encoded according to the ÖNACE specification [ÖNA]. In other words, IO Tables typically describe transactions on a nationality level, within a country or across a specific set of countries to convey import/export trends. (2) The second data category complements more detail by aggregating IO Tables' information with more granular/regional features, e.g., knowledge about individual firms within a region of a country. For this category we

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

used the “Sabina” dataset [Wir]: it contains the balance sheets of all the firms in Austria, including locations, cash flows, branches, and sector classification. By combining both datasets (see Sec. 4.5.1) we have the **Exchange Supply Chain (ESC)** dataset which was used during the design and Hermes‘ development. The network dataset features 172 nodes and 20,709 edges.

Users: Hermes target users are financial and economical experts involved in the development of regional policies and investment plans. Users are expected to be proficient in the statistical evaluation of tabular data, but are not expected to have any visual analysis or network analysis experience.

Tasks: Hermes aims at improving the understanding of national economic investment networks based on the integration of aggregated and granular data. Its goal is to provide information about the economic topology and support enhancements in goods and services’ main supply chains. After discussing with our domain experts collaborators, we could identified four sub-tasks: the exploration of the economic flow network existing between sectors (**T1**) and regions (**T2**), the evaluation of how sectors relate and impact each region (**T3**), and supply chain exploration (**T4**) (e.g., predominantly buying/selling regions). Identifying these four sub-tasks supports us to understand better the potential of our collaboration with the domain experts.

1.3.2 The Nested Model

Munzner‘s Nested Model characterizes different levels to design and implement visual solutions. The model is organized into four nested steps (see Figure 1.4). It starts from the examination of domain problem characterization, its challenges, and issues. Next, it maps these difficulties to digest data types and activities. In the third step, visual encodings and interaction techniques are picked. Last, the necessary algorithm design is implemented according to the steps above. This structure is genuinely valuable. It supports the design process and it allows for a constant evaluation during the solution development.

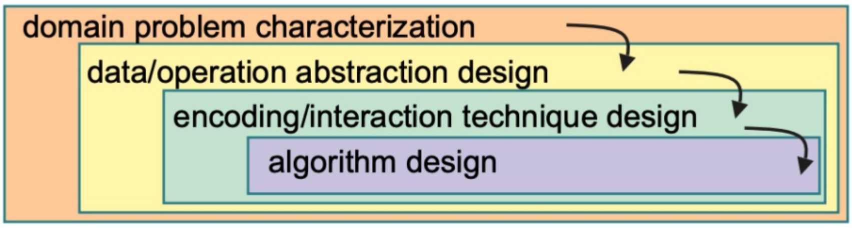

Figure 1.4: The Nested Model for visualization design and evaluation. Figure source: [Mun09a], ©2009 IEEE. Used with Permission

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

1.3.3 Research Workflow

By coupling features from the Design Triangle [MA14a] and the Nested Model [Mun09a], we derive the human-centered research approach that we followed during all solution development. Figure 1.5 shows a flow chart diagram of our research pipeline. First, we design mock-ups and propose possible workflows, which we evaluated with early feedback from our expert users. Depending on the evaluation, the existed mock-ups and workflows were adjusted and refined. The process repeated until we reached a satisfactory evaluation result. In the next step, we implemented a prototype according to the proposed solutions. Again, we evaluated the implemented prototype in a qualitative evaluation with the domain expert. We then derive a better understanding of the efficiency and suitability of the used techniques.

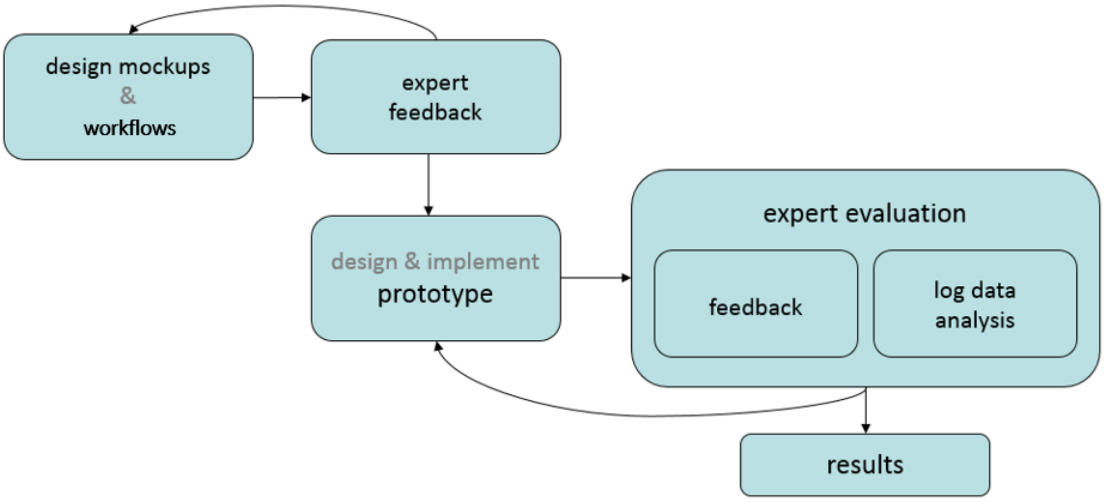

Figure 1.5: Human-centered research pipeline. During the prototype design and implementation, this approach allows a constant feedback loop between the experts and the development team until satisfactory results are reached. This loop was one of the main features always present during our research development.

We consider that the existing solutions (see Chapter 1.2.3) for event detection could be enhanced. These solutions do not exploit the full potential of VA in terms of interactive exploration and visualization. To improve the financial VA field and to fill the existing gaps, we propose a process pipeline (see Figure 1.6) that utilizes VA methods. From a literature review we derived several tasks of which we focus on (1) anomalous detection as well as (2) events monitoring. The anomalous detection task can also be expanded to the classification of such events. We trust that these two tasks should always be involved in operations that intent to analyse events.

Anomaly Event Detection The first task is the identification of unknown events and classification based on events that have already occurred. This process generates

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

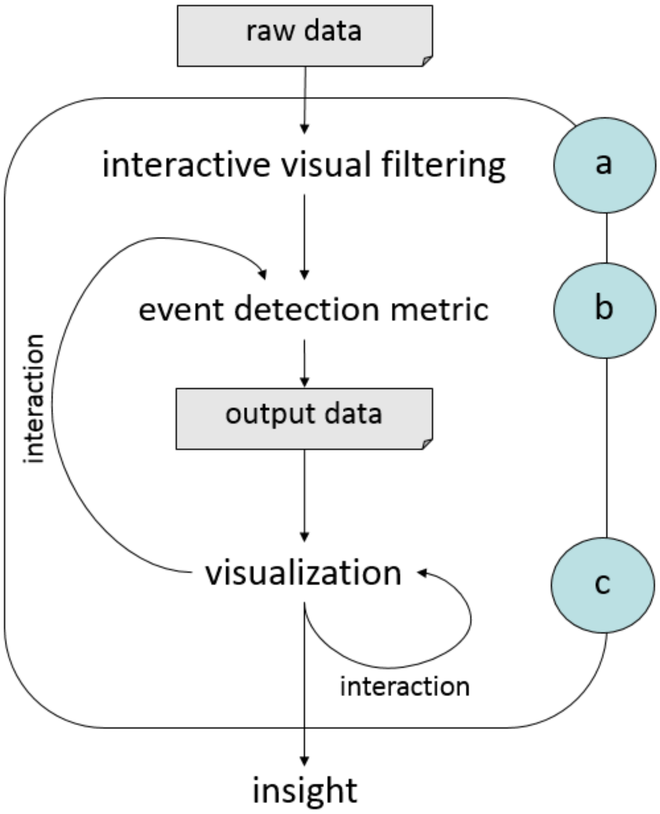

Figure 1.6: Financial VA generation process. Interactive visual filtering (a) allows for initial insights and filtering of the dataset. Metric for event detection (b) is the step where different risk management formulas can be applied and edited, generating "output data", which is used in (c). Visualization (c) is the phase where VA methods are employed to make the results from event detection metrics visually explorable. Interaction techniques and immediate visual feedback aid the user to fine-tune these metrics.

visual signatures for similar event cases. Those signatures can be later used, during monitoring, to identify potential hybrid attacks trough signature comparisons. This can be done automatically to some extend. However, events are dynamic and, therefore, are constantly changing. The visual comparison allows for appraising suspicious cases where automatic methods fail. This information can, then, be used to modify the detection metrics accordingly.

Event monitoring Costumers monitoring focuses on real-time monitoring of the data for an early identification or even prediction of anomalous behavior. To this end, we use the event detection mechanisms described above. Another feature that could be available is the enhancement of the "known events" database. This approach could also re-simulate past events to possible find non-classified events from the past. It could add

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

recently discovered variations of the same event to the "known events" dataset, which is used to aid real-time data monitoring. This process could make our monitoring system more flexible and accurate.

Our proposed process include the following steps: Figure 1.6 (a) we generate different interactive visualizations from the raw multivariate data which are used to filter the most interesting attributes as well as their value ranges. This step also allows for generating first hypotheses about attribute relationships.

In the next step (see Figure 1.6 (b)), metrics or AI techniques are applied. Fraud detection analysts need to be able to set and edit available metric formulas, and also create their own metrics. Previously generated signatures of different types of event are used to automatically identify anomalous cases. Subsequently, we present these results in an interactive visualization (see Figure 1.6 (c)). This visualization allows the analyst to dynamically modify the visual representation of data features to meet his/her interests and questions in mind. The visual comparison of event signatures helps to reason about differences and similarities. Insights from this exploration are then used to fine-tune the event detection metrics accordingly, closing the feedback loop (see arrows in Figure 1.6).

1.4 Objectives and Research Questions

This dissertation aims to answer the following main research question:

> **"How can VA techniques support event analysis by investigating time-oriented and multivariate data in different domains by accomplishing various tasks and users?"**

This primary research question can be split into the following sub-questions:

- **S1**: How can we address the particular characteristics of users, time-oriented and multivariate data, and tasks?
- **S2**: Can we derive an appropriate workflow for supporting different user groups and their tasks?
- **S3**: How can we generalize our contribution to other domains?

Complex time-oriented multivariate real-world datasets analysis can be substantially improved by visual analysis. Generic tools may aid initial data understanding and exploration. However, the potential of the VA solution is enhanced when designed and developed to target data-user-tasks. Linked interactive multi-views proved to support different domain users and tasks significantly.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Scientific Scope. This Ph.D. thesis is organized in two parts. In the first part, we present a state of the art contextualization that is later, in Part II, unified with our research results. In Part II, we present three main full papers. All articles cover VA solutions to event analysis involving similar data aspects (time-oriented and multivariate). Among the three presented articles, we supported two different domains, proving the flexibility of our methodology. More details about our articles and their connections are clarified in Chapter 1.5.

Financial operation management systems need to be in constant evaluation to avoid frauds and to provide risk management. Operations, such as a bank transaction and credit control, usually involve data with time-oriented and multivariate aspects. Due to its complex nature [Aig+11], time-oriented data, as well as multivariate data, require detailed exploration and analysis. Both aspects require flexibility and freedom of investigation and, by consequence, are targets of interest to the VA community.

Identify frauds in transaction datasets requires background knowledge and experiences of the domain experts. It is still unknown which and how VA features can enhance FFD. To aid investigators in this task, we improved interactivity with data through VA techniques. It would improve the knowledge about data and also introduce data exploration freedom to the different types of users. For example, one very useful feature in FFD is the possibility of saving and reusing fraud detection rules. Also, summarising results from data mining techniques could help search for fraudulent behavior within large amounts of data. Based on systematic state-of-the-art research, we designed and implemented a research prototype in order to aid FFD. We conducted a qualitative evaluation with domain experts to assess our envisioned VA techniques' appropriateness, effectiveness, and expensiveness.

1.5 Publications Summary

The developed approaches and its value to the scientific community have been published mainly in international scientific journals and our findings have been presented in specialized conferences.

Scientific Journals: Computer Graphics Forum (CGF), Transactions on Visualization and Computer Graphics (TVCG), Visual Informatics (Elsevier), Computer Graphics and Applications.

Specialized Conferences: IEEE VIS and EuroVis.

All publications involved collaborations with excellent scientific fellows from different renowned institutes. The collaborative aspect also supports the dissemination of our work results. Next, we present a cumulative journal publication list (see Section 1.5.1) and additional collaborative publication list (see Section 1.5.2). All publications within these lists were published in the last six years, most recently updated in December 2020.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

1.5.1 Cumulative Journal Publications

The following journal articles are reported in their entirety in Chapters 2, 3, and 4 as part of this cumulative dissertation:

[Lei+17] R. A. Leite, T. Gschwandtner, S. Miksch, S. Kriglstein, M. Pohl, E. Gstrein, and J. Kuntner. „EVA: Visual analytics to identify fraudulent events“. In: *IEEE transactions on visualization and computer graphics* 24.1 (2017), pp. 330–339

My Contributions: I was the lead author of this paper, proposing a VA approach to supporting fraudulent events investigation and fine-tuning a fraudulent event detection algorithm. During a collaborative work with domain experts from Erste Bank, I explored the financial fraud domain, its needs, and its challenges. To do so, I participated in frequent meetings with domain experts. Inspired by the existing literature, I designed different data visualization and interaction. I implemented prototypes and evaluated them cyclically with the domain experts. As lead author, I also conceptualized and wrote the evaluation sessions, collected and analyzed the outputs. Thus, I suggested an initial draft of the first paper version. Throughout the paper writing development, I was responsible for the consistency of the whole paper as well as for merging inputs of the co-authors. As the first author, I took care of the submission process and its review circles.

[A L+20] R. A. Leite, T. Gschwandtner, S. Miksch, E. Gstrein, and J. Kuntner. „NEVA: Visual analytics to identify fraudulent networks“. In: *Computer Graphics Forum*. Vol. 39. 6. Wiley Online Library. 2020, pp. 344–359

My Contributions: I was the lead author of this paper, proposing a VA solution focused on reducing the false-negative and false-positive fraudulent events detection in a network of customers. This work is a continuation of EVA [Lei+17], which focuses on the network aspect of a bank customer structure. The development of this work involved domain experts in a human-centered approach fashion. It had several implemented prototypes that were fine-tuned in collaboration with domain experts from the bank until it converged to a final form which was, then evaluated. Our collaborators coordinated the evaluation sessions from the bank, which were designed, executed, and analyzed by me. As the first author, I did an early version of the paper, which was subjected to improvement cycles involving the co-authors until we had a final version. I was also responsible for the paper submission and to improve it during the review circles.

[Lei+20b] R. A. Leite, A. Arleo, J. Sorger, T. Gschwandtner, and S. Miksch. „Hermes: Guidance-enriched Visual Analytics for economic network exploration“. In: *Visual Informatics* 4.4 (2020), pp. 11–22

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

My Contributions: I was the lead author of this paper, which proposes a guidance-enriched VA environment to explore complex economic network events. This work uncovers supply chains, regions' productivity, and sector-to-sector relationships. It was developed based on a real-world Austrian database: Sabina [Wir]. During this work, my primary duty was to design, prototype, and evaluate the approach. As lead author, I drafted the paper's initial version, which was then revised by the co-authors. I was also responsible for submitting the article and incorporating improvement suggestions during the review circles. I see this work as the sequence of NEVA [A L+20] when it comes to exploring network events.

1.5.2 Additional Publications

Additional publications where I am first author or co-author during the last six years:

[Lei+15b] R. A. Leite, T. Gschwandtner, S. Miksch, E. Gstrein, and J. Kuntner. „Visual analytics for fraud detection and monitoring". In: *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE. 2015, pp. 201–202

[Lei+16] R. A. Leite, T. Gschwandtner, S. Miksch, E. Gstrein, and J. Kuntner. „Visual Analytics for Fraud Detection: Focusing on Profile Analysis." In: *EuroVis (Posters)*. 2016, pp. 45–47

[Bög+17] M. Bögl, P. Filzmoser, T. Gschwandtner, T. Lammarsch, R. A. Leite, S. Miksch, and A. Rind. „Cycle Plot Revisited: Multivariate Outlier Detection Using a Distance-Based Abstraction". In: *Computer Graphics Forum*. Vol. 36. 3. Wiley Online Library. 2017, pp. 227–238

[Lei+18a] R. A. Leite, T. Gschwandtner, S. Miksch, E. Gstrein, and J. Kuntner. „Network Analysis for Financial Fraud Detection." In: *EuroVis (Posters)*. 2018, pp. 21–23

[Lei+18b] R. A. Leite, T. Gschwandtner, S. Miksch, E. Gstrein, and J. Kuntner. „Visual analytics for event detection: Focusing on fraud". In: *Visual Informatics (STAR)* 2.4 (2018), pp. 198–212

[Sal+19] S. M. Salisu, E. Mayr, V. A. Filipov, R. A. Leite, S. Miksch, and F. Windhager. „Shapes of Time: Visualizing Set Changes Over Time in Cultural Heritage Collections." In: *EuroVis (Posters)*. 2019, pp. 45–47

[May+19] E. Mayr, S. Salsiu, V. A. Filipov, G. Schreder, R. A. Leite, S. Miksch, and F. Windhager. „Visualizing biographical trajectories by historical artifacts: A case study based on the photography collection of Charles W. Cushman". In: *Paper accepted for publication at the Biographical Data in a Digital World Conference*. Vol. 2019. 2019

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

[**Arl+19**] A. Arleo, C. Tsigkanos, C. Jia, R. A. Leite, I. Murturi, M. Klaffenböck, S. Dustdar, M. Wimmer, S. Miksch, and J. Sorger. „Sabrina: Modeling and Visualization of Financial Data over Time with Incremental Domain Knowledge“. In: *2019 IEEE Visualization Conference (VIS)*. IEEE. 2019, pp. 51–55

[**Win+20**] F. Windhager, S. Salisu, R. A. Leite, V. Filipov, S. Miksch, G. Schreder, and E. Mayr. „Many Views Are Not Enough: Designing for Synoptic Insights in Cultural Collections“. In: *IEEE Computer Graphics and Applications* 40.3 (2020), pp. 58–71

[**Lei+20a**] R. A. Leite, V. Schetinger, D. Ceneda, B. Henz, and S. Miksch. „COVIs: Supporting Temporal Visual Analysis of Covid-19 Events Usable in Data-Driven Journalism“. In: *2020 IEEE Visualization Conference (VIS)*. IEEE. 2020, pp. 56–60

[**SLM21**] A. Scheidl, R. Leite, and S. Miksch. „VisMiFlow: Visual Analytics to Support Citizen Migration Understanding Over Time and Space“. In: IEEE, 2021

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Bibliography

[A L+20] R. A. Leite et al. „NEVA: Visual analytics to identify fraudulent networks". In: *Computer Graphics Forum*. Vol. 39. 6. Wiley Online Library. 2020, pp. 344–359.

[Aig+11] W. Aigner et al. *Visualization of time-oriented data*. Springer Science & Business Media, 2011.

[AK13] F. Atefeh and W. Khreich. „A survey of techniques for event detection in twitter". In: *Computational Intelligence* (2013).

[Arl+19] A. Arleo et al. „Sabrina: Modeling and Visualization of Financial Data over Time with Incremental Domain Knowledge". In: *2019 IEEE Visualization Conference (VIS)*. IEEE. 2019, pp. 51–55.

[ASS13] E. N. Argyriou, A. A. Sotiraki, and A. Symvonis. „Occupational fraud detection through visualization". In: *2013 IEEE, Intelligence and Security Informatics (ISI)*. IEEE. 2013, pp. 4–6.

[Bög+17] M. Bögl et al. „Cycle Plot Revisited: Multivariate Outlier Detection Using a Distance-Based Abstraction". In: *Computer Graphics Forum*. Vol. 36. 3. Wiley Online Library. 2017, pp. 227–238.

[CBK09] V. Chandola, A. Banerjee, and V. Kumar. „Anomaly detection: A survey". In: *ACM computing surveys (CSUR)* 41.3 (2009), pp. 1–58.

[CC08] J. D. Cryer and K.-S. Chan. *Time series analysis: with applications in R*. Springer Science & Business Media, 2008.

[Cha+07] R. Chang et al. „Wirevis: Visualization of categorical, time-varying data from financial transactions". In: *Visual Analytics Science and Technology. VAST. IEEE Symposium on*. IEEE. 2007, pp. 155–162.

[Cle93] W. S. Cleveland. *Visualizing data*. Hobart Press, 1993.

[DR15] W. N. Dilla and R. L. Raschke. „Data visualization for fraud detection: Practice implications and a call for future research". In: *International Journal of Accounting Information Systems* 16 (2015), pp. 1–22.

[Edw83] T. Edward R. „The Visual Display of Quantitative Information". In: (1983).

[Faw+98] T. Fawcett et al. „AI approaches to fraud detection and risk management." In: *AI Magazine* 19.2 (1998), p. 107.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

[Fun36] H. G. Funkhouser. „A note on a tenth century graph". In: *Osiris* 1 (1936), pp. 260–262.

[Fun37] H. G. Funkhouser. „Historical development of the graphical representation of statistical data". In: *Osiris* 3 (1937), pp. 269–404.

[Gau+13] D. Gauldie et al. „Louvain Clustering for Big Data Graph Visual Analytics". In: (2013).

[HLN09] M. L. Huang, J. Liang, and Q. V. Nguyen. „A visualization approach for frauds detection in financial market". In: *Information Visualisation, 13th International Conference*. IEEE. 2009, pp. 197–202.

[Ise+13] T. Isenberg et al. „A systematic review on the practice of evaluating visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2818–2827.

[Kei+08] D. Keim et al. „Visual Analytics: Scope and Challenges". In: *Visual Data Mining*. Ed. by S. J. Simoff, M. H. Böhlen, and A. Mazeika. Springer, 2008, pp. 76–90. DOI: 10.1007/978-3-540-71080-6_6.

[Kle13] G. Klein. *Seeing what others don't: The remarkable ways we gain insights*. PublicAffairs, 2013.

[KP15] S. Kriglstein and M. Pohl. „Choosing the Right Sample? Experiences of Selecting Participants for Visualization Evaluation". In: *EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (EuroRV3)*. Ed. by W. Aigner, P. Rosenthal, and C. Scheidegger. The Eurographics Association, 2015. DOI: 10.2312/eurorv3.20151146.

[Lei+15a] R. A. Leite et al. „Visual Analytics for Fraud Detection and Monitoring". In: *Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on*. IEEE. 2015, pp. 201–202.

[Lei+15b] R. A. Leite et al. „Visual analytics for fraud detection and monitoring". In: *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE. 2015, pp. 201–202.

[Lei+16] R. A. Leite et al. „Visual Analytics for Fraud Detection: Focusing on Profile Analysis." In: *EuroVis (Posters)*. 2016, pp. 45–47.

[Lei+17] R. A. Leite et al. „EVA: Visual analytics to identify fraudulent events". In: *IEEE transactions on visualization and computer graphics* 24.1 (2017), pp. 330–339.

[Lei+18a] R. A. Leite et al. „Network Analysis for Financial Fraud Detection." In: *EuroVis (Posters)*. 2018, pp. 21–23.

[Lei+18b] R. A. Leite et al. „Visual analytics for event detection: Focusing on fraud". In: *Visual Informatics (STAR)* 2.4 (2018), pp. 198–212.

[Lei+20a] R. A. Leite et al. „COVIs: Supporting Temporal Visual Analysis of Covid-19 Events Usable in Data-Driven Journalism". In: *2020 IEEE Visualization Conference (VIS)*. IEEE. 2020, pp. 56–60.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[Lei+20b] R. A. Leite et al. „Hermes: Guidance-enriched Visual Analytics for economic network exploration“. In: *Visual Informatics* 4.4 (2020), pp. 11–22.

[LF10] S. Lai and H.-C. Fu. „A polygon description based similarity measurement of stock market behavior“. In: *Expert Systems with Applications* 37.2 (2010), pp. 1113–1123.

[Lue10] J. Luell. „Employee fraud detection under real world conditions“. PhD thesis. University of Zurich, 2010.

[MA14a] S. Miksch and W. Aigner. „A matter of time: Applying a data–users–tasks design triangle to visual analytics of time-oriented data“. In: *Computers & Graphics* 38 (2014), pp. 286–290. DOI: 10.1016/j.cag.2013.11.002.

[MA14b] S. Miksch and W. Aigner. „A matter of time: Applying a data–users–tasks design triangle to visual analytics of time-oriented data“. In: *Computers & Graphics* 38 (2014), pp. 286–290.

[May+19] E. Mayr et al. „Visualizing biographical trajectories by historical artifacts: A case study based on the photography collection of Charles W. Cushman“. In: *Paper accepted for publication at the Biographical Data in a Digital World Conference*. Vol. 2019. 2019.

[Mun09a] T. Munzner. „A nested model for visualization design and validation“. In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009), pp. 921–928.

[Mun09b] T. Munzner. „A nested model for visualization design and validation“. In: *IEEE transactions on visualization and computer graphics* 15.6 (2009), pp. 921–928.

[ÖNA] ÖNACE. *ÖNACE 2008 - Struktur*. Accessed: 2019-07-20. http://www.statistik.at/KDBWeb/kdb_VersionAuswahl.do.

[RG13] D. Rosenberg and A. Grafton. *Cartographies of time: A history of the timeline*. Princeton Architectural Press, 2013.

[Sac+14] D. Sacha et al. „Knowledge generation model for visual analytics“. In: *Visualization and Computer Graphics, IEEE Transactions on* 20.12 (2014), pp. 1604–1613.

[Sak+10] C. Sakoda et al. „Visualization for Assisting Rule Definition Tasks of Credit Card Fraud Detection Systems“. In: *IIEEJ Image Electronics and Visual Computing Workshop*. 2010.

[Sal+19] S. M. Salisu et al. „Shapes of Time: Visualizing Set Changes Over Time in Cultural Heritage Collections.“ In: *EuroVis (Posters)*. 2019, pp. 45–47.

[ŠB10] A. Šilić and B. D. Bašić. „Visualization of text streams: A survey“. In: *Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2010, pp. 31–43.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

[Sch+11] M. Schaefer et al. *A Novel Explorative Visualization Tool for Financial Time Series Data Analysis*. Bibliothek der Universität Konstanz, 2011.

[Shn96] B. Shneiderman. „The eyes have it: A task by data type taxonomy for information visualizations“. In: *Proc. of IEEE Symposium on Visual Languages*. IEEE, 1996, pp. 336–343. DOI: 10.1109/VL.1996.545307.

[SLM21] A. Scheidl, R. Leite, and S. Miksch. „VisMiFlow: Visual Analytics to Support Citizen Migration Understanding Over Time and Space“. In: IEEE, 2021.

[TC05] J. J. Thomas and K. A. Cook. *Illuminating the path: The research and development agenda for visual analytics*. Tech. rep. Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2005.

[TG83] E. R. Tufte and P. R. Graves-Morris. *The visual display of quantitative information*. Vol. 2. 9. Graphics press Cheshire, CT, 1983.

[Til75] L. Tilling. „Early experimental graphs“. In: *The British Journal for the History of Science* 8.3 (1975), pp. 193–213.

[Tim+15] M. P. Timmer et al. „An Illustrated User Guide to the World Input–Output Database: the Case of Global Automotive Production“. In: vol. 23. 3. 2015, pp. 575–605.

[TK08] T. Tekušová and J. Kohlhammer. „Visual analysis and exploration of complex corporate shareholder networks“. In: *Electronic Imaging*. International Society for Optics and Photonics. 2008.

[Tuk77] J. W. Tukey. *Exploratory data analysis*. Vol. 2. Reading, Mass., 1977.

[Win+20] F. Windhager et al. „Many Views Are Not Enough: Designing for Synoptic Insights in Cultural Collections“. In: *IEEE Computer Graphics and Applications* 40.3 (2020), pp. 58–71.

[Wir] Wirtschaftsuniversität Wien. *Sabina - Info - Datenbanken*. Accessed: 2019-06-10. https://www.wu.ac.at/bibliothek/recherche/datenbanken/info/sabina/.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Part II

Papers

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

CHAPTER 2

EVA: Visual Analytics to Identify Fraudulent Events

This paper first appeared in: IEEE Transactions on Visualization and Computer Graphics 24.1: 330-339, 2017. ©2017 IEEE. Reused with Permission. (See [Lei+17])

Full List of Authors Besides Roger A. Leite, the following authors contributed and provided valuable input to this paper: Theresia Gschwandtner, Silvia Miksch, Simone Kriglstein, Margit Pohl, Erich Gstrein, and Johannes Kuntner.

2.1 Abstract

Financial institutions are interested in ensuring security and quality for their customers. Banks, for instance, need to identify and stop harmful transactions in a timely manner. In order to detect fraudulent operations, data mining techniques and customer profile analysis are commonly used. However, these approaches are not supported by Visual Analytics techniques yet. Visual Analytics techniques have potential to considerably enhance the knowledge discovery process and increase the detection and prediction accuracy of financial fraud detection systems. Thus, we propose EVA, a Visual Analytics approach for supporting fraud investigation, fine-tuning fraud detection algorithms, and thus, reducing false positive alarms.

2.2 Introduction

Event detection is an important task in many domains such as finding interesting changes in stock markets, spotting problems in health parameters, or detecting financial fraud. Analyzing these events in a temporal context allows the identification of insights such

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

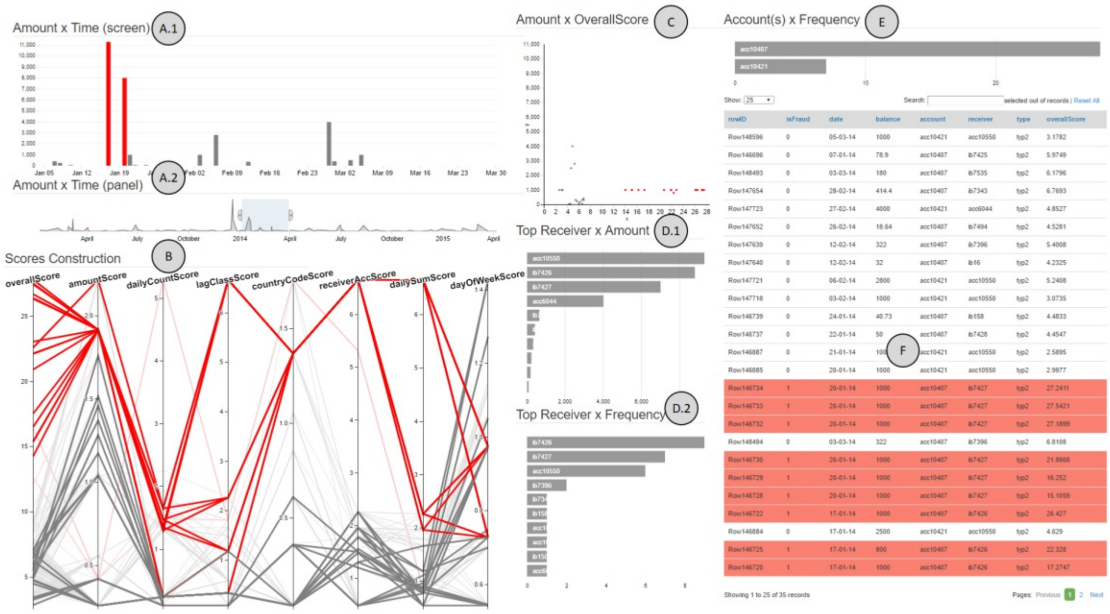

Figure 2.1: Screenshot of EVA (Event detection with Visual Analytics). (A.1, A.2) Temporal Views: a filter was applied in (A.2) to the period from January 2014 until April 2014. (B) Score Construction View: each line represents a transaction and its scores. (C) Amount vs Overall Score Scatterplot. (D.1, D.2) Ranks of accounts that received the highest amounts of money from the selected account and accounts that received the highest number of transactions from the selected account. (E) Accounts Selector: bars shows amount of transactions from each account. (F) Dynamic Table of raw transaction data. In all views, elements that represents suspicious data are highlighted in red.

as frequency, trends, and changes. Moreover, the investigation of outliers allows the analyst to identify risks, drastic changes, or rare occurrences. In this work we focus on the identification of anomalous events in the financial sector.

Financial institutions handle millions of transactions from clients per year. Although the majority part of these transactions being legitimate, a small number of them are criminal attempts, which may cause serious harm to customers or to the financial institutions themselves. Thus, the trustability of each transaction has to be assessed by the institution. However, due to the complex and multidimensional data at hand, financial fraud detection (FFD) is a difficult task.

The well renowned Oxford Dictionary defines fraud as "wrongful or criminal deception intended to result in financial or personal gain" [1]. Software environments handling sensitive data such as financial operation management systems, systems for insurance evaluation, or companies' internal control systems, need to be in constant evaluation to prevent fraud, to provide risk management, and, thus, to avoid serious consequences.

[1] http://www.oxforddictionaries.com/definition/english/fraud (accessed December 10, 2016)

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

All these scenarios deal with similar data with the aim to detect suspicious events and, thus, to identify frauds. For instance, the two tasks of monitoring bank transactions and credit control usually involve data with time-oriented and multivariate aspects. Due to its complex nature [Aig+11], time-oriented and multivariate data require sophisticated means for detailed analysis and exploration. By consequence, both are subjects of interest to the Visual Analytics (VA) community.

Besides its challenging nature, FFD has also a strong social and financial importance. For instance, fraudulent schemes such as 'money laundering', 'unauthorized transaction', or 'straw person' should be detected and fought as fast as possible by financial systems, since the negative economical and social impact increases with time. Thus, governments, banks, and other financial institutions that provide credit and money transaction services have a strong interest in improving operation monitoring and fraud detection.

Kielman et al. [KTM09] describe fraud detection as an open VA problem that requires visual exploration, discovery, and analysis. However, many of the current solutions involve mainly data mining techniques, while neglecting the potential of VA techniques to integrate human analysis into the process [Kei+08]. In this paper, we aim at closing this gap by presenting a VA approach for the investigation of suspicious financial transactions and fine-tuning of an existing automatic alert system. VA approaches may be utilized to identify different types of frauds. In this work, we focus on detecting "unauthorized transactions" within a financial institution. We designed our VA approach for FFD according to the nested model [Mun09] paying attention that our solution is flexible and extensible enough to be applied in similar domains with similar multivariate and time-oriented aspects. The main contributions are:

- In tight collaboration with domain experts we analyzed the real world problem of FFD and iteratively designed EVA, a VA approach to improve their current work flow;
- EVA interweaves well-known visualization techniques, which our domain experts are mostly familiar with, and automatic methods;
- To the best of our knowledge, we present the first VA approach based on a scoring system for FFD;
- We present our findings from an evaluation with three target users (not involved in the design process) and categorize the types of insights that could be gained with our prototype;
- We derived open challenges and possible future research directions in the field.

2.3 Related Work

There is a number of surveys that focus on fraud detection. In 2002, Bolton and Hand [Ric02] published a review about fraud detection approaches. They described the

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

available tools for statistical fraud detection and identified the most used technologies in four areas: credit card fraud, money laundering, telecommunication fraud, and computer intrusion. Kou et al. [Kou+04] presented a survey of techniques for identifying the same types of fraud as described in [Ric02]. The different approaches are broadly classified into two categories: misuse and anomaly detection. Both categories present techniques such as: outlier detection, neural networks, expert systems, model-based reasoning, data mining, state transition analysis, and information visualization. These works helped us to understand diverse fraud domains and how they are normally tackled. When looking on surveys of visual approaches for financial data, we identified FinanceVis [DML14] which is a browser tool including over 85 papers related to financial data visualization. FinanceVis was instrumental in analyzing how data that is similar to our data is usually visualized. Motivated by a lack of information, Ko et al. [Ko+16] presented a survey of approaches for exploring financial data. In this work, financial data experts were interviewed concerning their preferences of data sources, automated techniques, visualizations, and interaction methods.

When it comes to visual solutions to support FFD, Kirkland et al. [Kir+99] published one of the first works in fraud detection using visual techniques. In their work they combined Artificial Intelligence (AI), visualization, pattern recognition, and data mining to support regulatory analysis, alerts (fraud detection), and knowledge discovery. In our approach, we use a similar combination of techniques, but we also provide means for an interactive exploration of the visualized data.

WireVis's [Cha+07] main idea is to explore big amounts of transaction data using multiple coordinated views. In order to aid fraud detection, they highlight similarities between accounts based on keywords over time. Yet, WireVis does not support the detailed analysis of single accounts without clustering a set of accounts by their similar keywords usage. This is the most similar approach to EVA. However, instead of focusing on hierarchical analysis of keywords patterns within the transactions, EVA enables a broader and more flexible analysis. A deeper comparison with our approach is provided in Section 2.6.1.

A first financial data flow is presented by [SK95]. In this approach, data are aggregated in order to allow users to draw analytical conclusions and make transaction decisions. EventFlow [Mon+13] was designed to facilitate analysis, query, and data transformation of temporal event datasets. The goal of this work is to create aggregated data representations to track entities and the events related to them. When looking at approaches for event monitoring in general, Huang et al. [HLN09] presented a VA framework for stock market security. In order to reduce the number of false alarms produced by traditional AI techniques, this work presents a visualization approach combining a 3D tree map for market performance analysis and a node-link diagram for network analysis. Dilla et al. [DR15], presented the current needs in FFD. The authors presented a theoretical framework to predict when and how the investigators should apply VA techniques. They evaluated various visualization techniques and derived which visualizations support different cognitive processes. In addition, the authors also suggest future challenges in this research area and discuss the efficacy of interactive data visualization for fraud

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

detection, which we used as a starting point for our approach.

Carminati et al. [Car+14] presented a semi-supervised online banking fraud analysis and decision support based on profile generation and analysis. While this approach provides no visual support for fraud analysis, it is directly related to our approach since we are also focusing on profile analysis. However, we believe that VA methods have great potential to foster the investigation of the data and enable the analyst to better fine-tune the scoring system.

In the health domain, Rind et al. [Rin+13] conducted a survey study focusing on information visualization systems for exploring and querying electronic health records. Moreover, Wagner et al. [Wag+15] presented a systematic overview and categorization of malware visualization systems from a VA perspective. Both domains of these studies are similar to FFD, since they both involve multivariate and temporal aspects. However, the FFD domain demands for special consideration due to the complexity of the involved tasks (see Section 3.4.2).

2.4 Financial Fraud Detection

We developed our prototype called EVA (Event detection with Visual Analytics) in tight collaboration with a national bank institution [Ers] with the aim to improve and support their current FFD techniques.

In this section, we (1) describe the characteristics of transaction data, (2) discuss the complexity of the problem at hand, (3) present the currently used methodology for FFD at the bank, and (4) sketch EVA's scoring approach.

2.4.1 Transaction Data

We use an anonymized data set of real money transactions from our collaborating bank. This data set contains all transactions (e.g., payments, money transfers) executed or received by one of its customers within a given time period. Each transaction event is composed by several categorical, numerical, geospatial, and temporal dimensions. Some examples are: sender/receiver, amount of money, location, and time of execution. The combination of these different aspects of data results in complex analysis scenarios that require the combination of different techniques in order to be tackled. More details concerning the data set used during development and evaluation are given in Section 4.6.

2.4.2 Problem Complexity

Automated FFD techniques are suited for well-defined problems and scenarios where the investigator knows exactly which patterns he/she is examining. However, the majority of fraudulent cases are not easily predictable by common rules and require some human investigation. Consequently, new methods such as VA are needed for these ill-defined

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

problems. Besides the complexity that comes with the multivariate data set, there are several additional aspects that add up to the complexity of FFD.

Scalability. Financial institutions execute hundreds of thousands of transactions per day. To validate the veracity of all these transactions requires visually and analytically scalable solutions [Lei+15].

Context complexity. To better understand frauds, we need to consider the motivations that guide this criminal act. It is known that geopolitical, social, and economical contexts influence this criminal behavior [Jai11]. Considering the ever changing local and global scenarios, FFD techniques need to be adapted frequently.

Frequent Changes. Not only there are many different types of frauds, but new ones are constantly being created and old ones are constantly being adapted in order to hide from current detection mechanisms.

False Positives. For each transaction that is flagged as suspicious by the automatic system, an investigator has to decide if the accusation of fraudulent behavior is correct, or not. Depending on the fraudulent classification (i.g., in case of money laundering suspicion), the owner account is then sued. To bring the accusation to court, involves professionals and costs a lot of money. This means that as the levels of positive alarms increase, the bank wastes money and, also, looses customers. Besides, even if identified during the process, false positive alarms overloads the investigators and waste their time of analysis.

False Negatives. Frauds that are neither detected by the automatic scoring system nor by investigators produce a financial damage to the bank and impact its clients‘ safety. They also impact the trustworthiness of the institution. Moreover, false negatives overlook actual recurrent frauds and, by consequence, result in fraudulent harm [Lue10]. In other words, in order to be more helpful than harmful, the solutions need to be precise in estimating possible threats.

Fraud Classification. Fraudulent techniques are constantly being updated and reinvented. The definition of a set of features that classify fraud techniques is a difficult task which increases with the amount and complexity of data dimensions.

Time-Oriented Analysis. FFD not only requires the identification of temporal outliers, but also of periodic behavior (e.g., disguising fraudulent transactions as monthly bill payments). If well planned, frauds can avoid automatic algorithms detection. Thus, synchronous and asynchronous temporal aspects should be observed during analysis. However, due to its complexity [Aig+11], there are many aspects of temporal data that need to be analyzed efficiently.

2.4.3 Methodology for FFD

In this subsection we give an overview of the methodology that is used for FFD by our collaborating bank. Since we are using real data that is quite sensitive, we need to respect

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

privacy and security regulations. Thus, we are not allowed to get into details about the actual algorithms. However, we roughly sketch the four phases of the FFD methodology applied: Profile Generation System, Scoring System, Results Interpretation, and Fraud Validation.

Profile Generation System. For each customer account the automatic system for FFD generates profiles based on this account‘s transaction history (see Figure 3.1 A.1, A.2, and A.3). A single account can have several profiles which reflect different aspects, for instance, separate sender and receiver profiles for one account. The result of this profile construction is then used for further classification. Profile generation is not a phase that is sequentially linked with the other phases. It has its own rules of execution. The bank can define a period of time for when it has to be executed (every week), or after a certain amount of events (after 100 transactions).

Scoring System. The system compares each of the incoming transactions (see Figure 3.1 B.1) with the corresponding customer‘s profile. To this end, it uses metrics to compute several different scores that are summarized in one overall score (see Figure 3.1 B.2); a single float number that represents how suspicious a transaction is. For example, each time a customer makes a new transaction, the FFD system automatically compares this transaction with the customer‘s profile of past transactions in order to compute a score that flags this transaction as either suspicious (possible fraud) or not suspicious. The higher the score, the more suspicious is the transaction (i.e., different aspects of the transaction that influence the score).

Results Interpretation. After calculating the scores, the non-automatic part of the investigation takes place. In this phase, investigators analyze multiple transactions simultaneously, due to time constraints. Transactions whose scores exceed a given threshold are further filtered by predefined rules. For example, all transactions below 20 euros are excluded from the list of suspicious transactions, since the amount is too low. The remaining transactions are then manually explored with the help of spreadsheet tools. During this exploration, investigators use their personal experience to decide whether a transaction should be considered fraudulent or not (see Figure 3.1 D).

Fraud Validation. Once investigators have decided that a suspicious transaction is possibly fraudulent, they call the account owner to ask him/her about the transaction‘s veracity. The bank stops the transaction in case the account owner did not authorize it.

We incorporated a VA component into the described work flow to tackle the complexity of fraud analyses (compare Section 3.4.2). The new process is illustrated in Figure 3.1.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

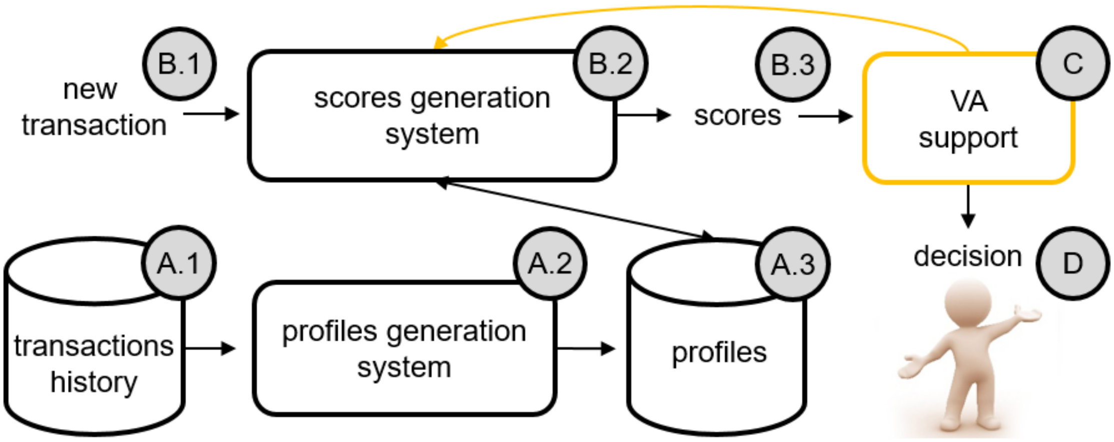

Figure 2.2: The transaction evaluation system. The newly added interactive VA approach for investigating suspicious behavior and for evaluating the scoring system is highlighted in orange.

Types of Fraud

Different scores can be constructed for representing different types of frauds. There are more definitions of fraud types and their subtypes than the ones described in this section. However, we opt for describing the ones that mattered most to our collaborators.

Money Laundering. The main goal of money laundering is to transform profit gained from crime and corruption into 'legal' money. Usually, this type of fraud is composed of a number of events involving a network of accounts. Thus, the computation and analysis of scores needs to include network analysis for detecting this type of fraud.

Unauthorized Transaction. This type of fraud involves transactions from the account of a customer of a financial institution made by a non-authorized user with the aim of financial profit. This fraud is usually detected by profile analysis, i.e., comparing this transactions to other transactions usually done by this customer. Uncommon transactions receive high scores and need to be further investigated.

Embezzlement. In this type of fraud a criminal person misappropriates the money entrusted to him/her. This fraud may happen in the public or private sector and it is usually considered an internal fraud. To detect this fraudulent behaviour, scores need to consider transaction flows. This is usually done based on records from management software (e.g., log-files).

Straw Person. This type of fraud is sometimes related to money laundering. It describes a person A who receives money instead of person B, because B is not legally allowed to receive this money. In order to detect this type of fraud, scores need to consider customer profiles and detect outliers.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

During the design and development of EVA, we focused on detecting "unauthorized transactions". Since EVA uses scores for decision support, an extension of these scores to detect other types of frauds would be an appropriate way to perform other types of fraud analyses.

2.4.4 EVA's Scoring Approach

Our profile-based algorithm is a self-adaptive, histogram-based approach according to the (mandatory) guidelines from the European Banking Authority (EBA) and monitors the behavior of internet originated payments. The proposed algorithm computes individual customer profiles, which are created on basis of historical transactions. These profiles are used to score new transactions in real-time. Thus, depending on the relative deviation of the score from the profile's standard range, the payment might be classified as suspicious.

Due to privacy and security regulations of our collaboration partners of the bank, we are not allowed to describe our profile-based algorithm in detail. However, our approach is comparable with Carminati et al. [Car+14], who generate customer profiles in a semi-supervised way and provide different kinds of statistical analysis. As a result, their approach correctly ranks complex transactions as suspicious.

We evaluated EVA's profile-based algorithm on a representative sample of internet-based transactions consisting of 13 million payments ranging over a period of 15 months (1.1.2015 - 31.3.2016). To create the customer profiles the transactions of the first 12 months were taken (year 2015, 11.9 mill). For scoring - and consequently for evaluating - the remaining transactions of 2016 (about 1.1 million single transactions) were used. 24 transactions - out of this 1.1 mill - were flagged as confirmed frauds. Furthermore, as a reference system, an amount-threshold-based strategy was implemented, thus simulating the detection rules previously used.

Our profile-based approach documented a good performance and outperformed the threshold-based previously used strategy by far. For example: Taking the current number of confirmed fraudulent transactions identified by the threshold-based strategy as the constraint to be met, our approach found 500% more confirmed fraudulent transactions, and thus, preventing 86% of all potential amount losses. In a statistical analysis, applying Receiver Operating Characteristic (ROC) curve and calculating the Area Under the Curve (AUC) on both approaches, the profile-based approach produced an AUC of 0.944 while the threshold-based approach demonstrated less efficiency with an AUC of 0.78.

2.5 Design and Implementation

In the design phase of EVA we collaborated with two domain experts from the bank institution (referred as "collaborators"). Following the design triangle [MA14], to generate interactive VA methods we designed EVA with respect to the data, users, and tasks at hand.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Data. Financial transaction events constitute multivariate and time-oriented data which include details about the transactions such as amount, time, receiver, etc.

Users. Investigators from financial institutions that investigate and validate transactions alerts.

Tasks. The overall tasks are fraud detection by means of profile analysis. This task includes the reduction of false negative and false positive alarms, history comparison, as well as the manual investigation of suspicious transactions.

2.5.1 Requirements

When looking at currently used FFD solutions, there are still many opportunities for improvements. Instead of running queries in spreadsheets and judging alarms by a single overall score value, we propose EVA to support investigators during their decision-making process. From the study of related work and in collaboration with our project partners, we derived the follow requirements:

R1: Visual Support for Scoring System. Considering the constantly changing fraudulent behaviour, the scoring system and the profiling systems should be in constant evaluation. They should be frequently updated in order to stay effective. In the current system, investigators are not able to explore which transaction features and which sub-scores influenced the overall score to what extent. This information would be beneficial for understanding the construction of scores and deciding if the algorithm needs to be adapted. Moreover, investigators should be able to compare single transactions and their scores with the client‘s history of transactions.

R2: Account Comparison. Another important task in order to better understand suspicious events is to analyze the relationship between two accounts (i.e., their money exchanging behaviour). However, currently, this task is not supported besides the manual analysis of the two separate accounts by means of spreadsheets. Our solution needs to support the analysis of money exchange relationships of different accounts to enable the user to analyze and detect fraudulent collaborators.

R3: Reasoning About Potential Frauds. During the fraud validation phase (see Section 2.4.3), the investigator has to decide if a transaction flagged as suspicious is going to be confirmed as being fraud or not. To aid this task, our system needs to provide visual means to support the investigation of the automatically computed results. The system needs to allow for visually analyzing flagged transactions in contrast with non-flagged transactions, and thus, support the identification of false-positively flagged transactions.

R4: Identification of Hidden Frauds. Due to the data complexity (see Section 3.4.2), automatic methods such as the one used in our approach are not fully

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

accurate. This can lead investigators to overlook fraudulent transactions that were not detected by the automatic system. In order to better support this task, our solution needs to make similarities between flagged and non-flagged transactions visible during the validation phase (see Section 2.4.3). Thus, the system needs to facilitate the identification of false-negative frauds.

2.5.2 Event Detection with Visual Analytics (EVA)

Following a user-centric iterative design process [Mun09] we had regular meetings with our collaborators (about two hours each other month for one and a half years). We discussed the data, users, and tasks at hand in order to gain a thorough understanding of the problem and we designed a number of prototypes, ranging from low-fidelity mock-ups to interactive prototypes. Some design ideas we had to discard while others were iteractively refined and integrated into the final prototype. EVA is composed of six views displaying different aspects of the data (see Figure 2.1). All views are connected via brushing and linking (i.e., multiple coordinated views). EVA was developed as a web application by using Angular and D3.js technologies.

After some discussions with our collaborators, we opted for simple and well-known visualizations they are mostly familiar with, such as bar charts, line charts, and scatter-plots [Cle84; DR15]. The goal was to keep the learning curve for investigators as low as possible, and thus, faster acceptance of the system. However, since investigators are used to exploring data by using spreadsheets, we also provide a table representation in Figure 2.4, **F** which allows for assessing all details of the underlying data.

The visualization techniques were chosen with respect to the suitability of their visual attributes (e.g., element position, length, angle, color) to effectively and accurately encode the data types at hand [Mac86]. In particular, we chose different visual encodings to achieve the best possible balance between distinguishability, separability, and pop-out of important information.

A.1, A.2: Temporal Views. Both views represent the temporal dimension of transactions. In both views, time is laid out on the x-axis, while the y-axis represents the total amount of money transacted per day. Thus, in **A.1** (see Figure 2.4), each bar represents a day. Days that contain at least one suspicious transaction are highlighted in red. **A.2** (see Figure 2.4) serves as an overview visualization of the inspected time period and as an interactive temporal filter. View **A.1** is tightly linked to view **A.2**. When brushing an interval in **A.2**, view **A.1** zooms in or zooms out to this interval. Selected periods of time are propagated to the other views for the analysis of other data characteristics. In particular, **A.1** shows a more detailed temporal representation of the selected period, while A.2 preserves an overview of the whole time interval.

We opted for a line chart in **A.2** because this representation of time series data is well-known to the investigators. Moreover, our data at hand contains daily sums of money transacted, which are usually quite stable. However, there are days with unusual

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

high amounts of money transacted and these are the interesting days to identify and investigate, since high amounts of money may indicate frauds. Line charts allow for efficiently identifying such peaks even in very long time series. Alternative representations of time series data, such as calendar heatmaps would have been feasible too, but they bring some weaknesses about. These weaknesses include requiring more space, a more complex selection of temporal intervals, and a less intuitive visual encoding of the daily sum of money by color. For a more fine grained selection of days, we provide selectable bars in **A.1**. View **A.1** also represents the daily amount of money transacted on a temporal x-axis. However, daily amounts are represented by bars. In contrast to View **A.2** we decided for using bars in View **A.1** since investigators need to select suspicious days to investigate them in detail. For such a selection bars offer self-contained bodies each representing a single day, which can be more easily selected than regions on a line chart. Moreover, bar charts foster the accurate perception of the data by using bar length to encode quantitative information, which is accurately perceived [Mac86] and thus, presents a suitable visual encoding for this type of data. Discussions with our collaborators also showed that selecting one or multiple bars in a bar chart presented a simple and intuitive way of filtering the data.

B: Score Construction View. We use parallel coordinates to present a visual history of transactions where each line represents a transaction of the selected account (see Figure 2.4, view **B**). Transactions whose overall score exceeds a given threshold are considered suspicious and are highlighted in red. Besides the "overallScore" axis at the very left (which was an explicit request of our collaborators), all axes represent sub-scores computed by the automatic system that are used for constructing the overall score. Thus, this view supports investigators to understand how overall scores were constructed by the automatic system and how the score of each transaction fits into the overall scoring scope. This view supports filtering by brushing any set of axis and these filters are reflected in all the other views. Moreover, selections in any other view also filter the transactions displayed in this view. The Score Construction view **B** highlights the selected data while graying out other transactions. This feature allows investigators to keep the context of filtered transactions.

Parallel coordinates are well suited to represent multiple dimensions side by side, which makes them a rational choice for representing the different sub-scores that contribute to the overall suspiciousness score of transactions. Although the investigators were not familiar with parallel coordinates we still decided to use them for various reasons. We needed to provide a visualization that enables investigators to identify sub-scores with strong influences on the overall score. They also need to identify groups of transactions with similar sub-scores in order to better understand fraudulent patterns. In a previous version of the prototype, we used a scatterplot matrix to show these relations. However, this visualization technique confused our collaborators, while the parallel coordinates were well perceived (see Section 4.6). The sactterplot matrix provided too many scatterplots that failed to give an effective overview and only allowed for analysis of pairwise relations between sub-scores. All sub-scores of one transaction could only be related by brushing

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

and linking the dots in the different scatterplots, while parallel coordinates represent transactions by lines and the connection of all scores of such a transaction can easily be spotted. Thus, parallel coordinates facilitate comparing and relating these scores when reasoning why some transactions scored high or low. In addition, representing transactions by lines instead of dots in separate scatterplots also facilitates the identification of groups of transactions with similar patterns.

C: Amount vs Overall Score Scatterplot. In this view each dot represents a transaction. The overall score is encoded on the x-axis, while the amount of money exchanged is encoded on the y-axis. Thus, clusters of dots represent transactions with similar characteristics in these two dimensions, while outliers indicate uncommon transactions. Since investigators are interested in identifying outliers in contrast to clusters of normal transactions, overplotting in regions of normal transactions does not present a problem.

We decided for a scatterplot since it most efficiently encodes the relations between two variables. Using a scatterplot allows investigators to select a group of transactions (dots) according to the amount of money transferred and their overall suspiciousness score by area brushing. This supports the analysis of the relation of two of the most important dimensions for fraud detection: investigators emphasized that the amount of money is always a good place to start the investigation since small amounts of money are not of interest to them; combining this information with the overall score of a transaction facilitates the identification of cases that require further investigation. On the other hand, also transactions with high amounts of money that did not score very high are easily identified as outliers in this scatterplot and may hint at false negative cases.

D.1, D.2: Ranks. For analyzing money exchange relationships among clients, we provide two bar charts. These visualizations are utilized to represent who received money from transactions of the selected account. View **D.1** shows the rank of the top 10 receivers that received the biggest amount of money from the currently selected account, while the bar length encodes the sum of money received (see Figure 2.4). In **D.2**, we display the top 10 accounts which received money most frequently from the selected account, and thus, the bar length encodes the number of transactions received. Both types of information are important since frequently transferring money to the same receiver can hint at a fraudulent pattern, as well as transferring high amounts of money to one receiver. Investigators can select different bars in these two views to filter the data in all other views to show only transactions to the selected receivers. This way investigators can detect temporal patterns (e.g., frequent transactions to a specific receiver), analyze the history of transactions to this receiver, how they were ranked by the automatic scoring system, and drill-down into money exchange details by means of the Dynamic Table **F**. We decided for using bar charts to represent this information since they give a good overview of the ranking relationship of different receivers (i.e., very frequent receivers are emphasized by both, position at the top of the chart and bar length). Moreover, bars again allow for easy selection of interesting receivers for filtering and further investigation.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

E: Accounts Selector. When investigating more than one account, this view facilitates comparison and switching between accounts. The bar length represents the amount of transactions that each account executed, which already facilitates the selection of accounts of interest. By selecting a bar, investigators filter the other views to show only data of the selected account. This functionality can be used for comparing a small group of accounts in more detail.

F: Dynamic Table. Currently, investigators are used to apply queries within spreadsheets in order to find insights. Besides providing a good amount of details, tables hinder pattern recognition and scale badly. However, tables are known and appreciated by investigators and thus, we provide an interactive table view in addition to the other views. Each row represents one transaction and each column one of its dimensions. Filters and selections in other views are automatically reflected by the table view and the other way around. Moreover, it is possible to sort rows by column values and to execute manual search queries.

Multiple Connected Data Perspectives. Since transaction logs are composed of multiple heterogeneous dimensions that need to be analyzed in relation to each other, EVA provides multiple perspectives on the data in multiple connected views. This set of views presents a variety of abstraction levels of the same subset of the data. In all views that represent transactions, we chose a colorblind-safe color encoding [HB03] to indicate transactions flagged as suspicious. Using the color red makes these suspicious transactions stand out immediately.

2.6 Evaluation

To assess the usefulness of EVA, we conducted a qualitative evaluation which aimed to address the following research questions (RQ):

RQ1: Comparison. What are the advantages and disadvantages of EVA compared to the tools which users usually use?

RQ2: Insights. What kind of insights can be generated with EVA?

RQ3: Improvements. Do users miss any features or have suggestions for improvement?

We decided for a qualitative study because it allowed us to get users' feedback and to understand insights they gained while using EVA.

Sample. We recruited three target users of EVA, i.e. FFD investigators, from our collaborating bank, who were not involved in the design process and have never seen our prototype before. Although the number of participants was low, qualitative evaluation

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

studies are still useful to understand if the approach is useful for domain experts and if it fits their workflow [Ise+13; KP15]. All three male participants had basic knowledge of working with visualizations. They usually use visualizations for presentation purposes to show the key message and the structure of the data. However, one participant also noted that he uses visualizations for exploration tasks (e.g., to analyze algorithms via heatmaps). For fraud detection tasks, they primarily use rule-based management systems which provide mainly spreadsheet representations including the automatic generated scores for each transaction.

Dataset and Tasks. We used an anonymized real world dataset from our collaborators covering an interval from January 2013 to April 2015. The dataset consists of 413 different accounts with a total of 1,128,147 transactions of different types (e.g., netbanking transactions). These tasks are structured according to the analytical task taxonomy by Andrienko and Andrienko [AA06], distinguishing elementary and synoptic tasks. In order to evaluate our solutions with respect to our requirements, we have defined a list of typical tasks together with two collaborating domain experts. Each task was designed with a specific focus on one or more requirements (see Figure 2.3). Requirements such as interactivity **(R2)**, data conservancy **(R3)**, and visual scalability **(R5)** were considered in all tasks.

Task 1: Explore the top three frequent receivers from the account acc10407 during the period of January 2014 to April 2014.

Task 2: Explore the transactions of account acc10421 and find the reason about the scoring of all transactions that happened on day(s) where fraud(s) were detected.

Task 3: Analyze two fraudulent accounts (acc10407 and acc10421) with respect to their similarities and differences in their fraudulent behavior.

	R1	R2	R3	R4
Task 1		•	•	
Task 2	•		•	•
Task 3		•	•	•

Figure 2.3: This table shows the relation between tasks and requirements in our evaluation.

Procedure. The study took place in a quiet meeting room at the bank‘s head office. In addition to the respective participants, one test moderator, one observer for taking notes, and one developer as contact person for technical questions were present in the room. Furthermore, audio recording and screen capturing software was used. The test session began with a short introduction of the goal and the schedule of the study. Next, EVA was presented and participants had the possibility to ask questions to clarify any issues.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

We then conducted a semi-structured interview with the participants in order to learn about their experience regarding visualizations and which tools they usually use to solve their fraud detection tasks. After the interview, the participants were asked to interact with EVA in order to fulfill the three tasks outlined above. While the participants interacted with the prototype, they were encouraged to think aloud. After they finished the tasks, again a semi-structured interview was conducted. They were asked about their impressions of EVA, if they missed anything in particular, to compare the prototype with the tools they would typically use for fraud detection, and if there were any further tasks which they would like to solve with this kind of VA tool.

Data Analysis. The collected qualitative data (observation and interview notes as well as the audio and video recordings) were analyzed in order to find out what works well, what needs further improvement, and what are possible missing features (cf. research questions RQ1 and RQ3). However, we were also interested in which kinds of insights they gained with EVA while they solved the tasks (cf. RQ2). EVA supports processes of exploration and sensemaking. There are two well-known approaches explaining sensemaking with visualization - the model by Pirolli and Card [PC05] and Klein‘s sensemaking model (see also [KMH06a; KMH06b]). The model by Pirolli and Card has been criticized because it applies only to a very narrow range of activities of intelligence analysis [PSM12], while Klein‘s model is much broader. Therefore, Klein‘s categories were chosen for this analysis. Thus, we adapted the five categories from Klein [Kle13] for gaining insights:

Connection. These insights result from a connection between two or more events which provides new information. For example, two visualizations present the same data set from different viewpoints. The combination of these visualizations allows the viewer to get additional detail information about the data.

Coincidence. Coincidence insights result from events which seem related but do not have an obvious connection. In contrast to the connection insights a coincidence insight results from repetition and not from detail information. For example, if data points have the same value in the visualization then this can be a result from a specific event.

Curiosity. These insights differ in one way from the coincidence insights: it results from a single event. For example, a data point with a specific value in a visualization arouses the interest of the viewer.

Contradiction. Such insights often occur if there is discrepancy between events which causes doubts. For example, a data point in the visualization has an unrealistic value.

Creative Desperation. These insights result from events which tend to be a dead-end and require finding new ways. For example, if it is not possible to get relevant information with a specific type of visualization then another visualization type might be helpful.

Based on these categories, the observation notes as well as audio and video recordings were coded and categorized.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

2.6.1 Results

All participants solved all tasks. The average duration needed for the tasks was about 18 minutes. The interview sessions (before and after the participants solved the tasks) took about 40 minutes in total. Next, we will present and discuss the results according to our research questions.

RQ1: Comparison

The investigators stated that they typically use visualizations for presentation tasks which they typically generate with Microsoft Office tools (e.g., Excel and PowerPoint) [Mica; Micb]. Therefore, they argued that it is difficult to compare EVA with these tools. The challenge in using these tools is to find the interesting hot spots. All three investigators agreed that a powerful visual tool for exploration tasks would be helpful for browsing the data, and for gaining insights which they were not even looking for, or for giving them hints to look at specific parts in the data set more closely. They highlighted that EVA was very intuitive (e.g., color-coding), easy to use, more dynamic than the tools which they usually use and that it provides a good overview and quick access to detail information. For example, one expert mentioned: *"I liked it because it is very interactive and you can browse the data, even if you don't know what you are looking for, and find new insights"*. Furthermore, the usage of EVA led to a positive attitude towards using VA approaches in the future. For example, one expert noted: *"Here we could see what is possible [...] So we can rethink what we can offer to the bank."*.

Since, the investigators did not know any visual approaches to support FFD, they could not compare EVA to actual VA approaches. However, if we compare EVA, for instance, to WireVis [Cha+07], a state of the art VA approach focusing on FFD, EVA takes more aspects into account in order to identify fraudulent behaviour. While WireVis is used to analyze keywords used in transaction descriptions with a focus on detecting money laundering, EVA is aimed at detecting unauthorized transactions by analyzing a variety of objective aspects about transactions (e.g., amount, date and time, frequency, etc.). Thus, EVA presents a broader approach that is in line with existing FFD mining techniques. Moreover, EVA allows for a deeper exploration of multiple aspects of transactions as well as reasoning about how they influenced the automatically generated scores for fraud detection. Instead of analyzing keywords, our FFD approach constructs individual profiles for each account and computes suspiciousness scores that indicate how unusual the transaction is, given the history of this account. This supports the detection of different types of fraudulent transactions which would not be possible from keyword analysis alone.

RQ2: Insights

In total we found 77 insights. Most of these insights were connection insights (53.2%). Coincidence insights contributed 26%. Curiosity insights (6.5%) and contradiction insights (9.1%) played a marginal role. Creative desperation insights also only showed up in 5.2%

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

of the cases. Most insights (35 insights) were found for the **Task 3**, followed by **Task 2** with 24 insights and **Task 1** with 18 insights. The number of found insights correlated with increasing task complexity. For example, **Task 1** focused on the identification of specific values in order to offer the investigators an easy start with EVA whereas **Task 3** allowed for more data exploration since it required to analyze and compare two accounts in order to find their similarities and differences. Next, we will discuss each insight category in more detail.

Connection. In total, 41 connection insights were found and two types of connection insights were identified. Connection insights from the first type resulted from a connection between the different views of EVA. For example, one expert compared the Dynamic Table view (see Figure 2.1, F) with the Score Construction view (see Figure 2.1, B) to detect connections between the amount and the scores of the suspicious transactions. Connection insights from the second type, resulted from a connection between the different variables. For example, one expert noted: *"country code and daily count are suspicious since unusual many transactions were made from this foreign country"*. In total, they derived slightly more connection insights from the views (53.7%) than from the variables (46.3%). These results show that the different views helped investigators to analyze the data from different viewpoints. However, also the comparison of the different variables played a role in finding insights.

Coincidence. From the 77 found insights 20 were coincidence insights. These insights resulted from comparing values of the same variables. For example, one expert noted: *"the chance is high that these both transactions are also fraudulent since the receiver has already a confirmed fraudulent transaction"*. Or another investigator mentioned: *"when you detect one fraud in this case, you can detect all frauds because they were all made by only one person"*. Although we found more connection insights than coincidence insights during **Task 1** (15 versus 2 insights) and during **Task 2** (13 versus 4 insights), slightly more coincidence insights than connection insights were gained during **Task 3** (13 versus 14 insights). It seems that the differences between **Task 3** and the other two tasks arose from the comparison of the two accounts in **Task 3**. For example, we observed that the investigators compared the values of the variables separately for each account and next compared these between the accounts to detect similar behaviour.

Curiosity. We found 5 curiosity insights. These insights resulted from investigators' observations which stimulated their interest to explore the data further. For example, although one expert had already detected fraud cases with the bar chart visualization in the Time Panel (see Figure 2.1, A.1), he interacted further with the time slider from the area visualization (see Figure 2.1, A.2) in order to detect further possible cases. Curiosity insights only occurred during the last two tasks which were more exploratory in nature than the first one (**Task 2**: 3 insights and **Task 3**: 2 insights).

Contradiction. In total, 7 insights were contradiction insights. The contradiction insights arose from conflicts and doubts in their own observations but also in EVA. For example, one expert explained his decision not only to select the suspicious cases automatically marked from EVA: *"I would also select the surrounding in the scatterplot - because*

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

maybe the system did not detect all fraudulent cases". Most contradiction insights were found during **Task 3** (5 insights). It seems that comparing two accounts added to the complexity of Task 3. This showed that the participants had less confidence in their own observations. Thus, we plan to find solutions to minimize users' doubts in the future, for example, by directly highlighting the differences between accounts.

Creative Desperation. Only 4 creative desperation insights were found. These insights resulted from revising their own, previously phrased interpretations or from finding alternative ways when the desired interaction or view was not available. For example, one expert assumed that the interaction technique linking and brushing is not possible between the Dynamic Table view (see Figure 2.1, F) and the Score Construction view (see Figure 2.1, B) since he saw no changes between the two entries in the parallel coordinates after he selected them in the table. After he tried a third entry, he realized that the first two entries had the same values and hence there were no visual differences in view B.

RQ3: Improvements

The investigators made useful suggestions for possible new features and improvements. One suggestion was to expand the filter and selection functionality. For example, all investigators noted that filter options especially for the table representation (e.g., only to show transactions with a certain amount) would be helpful. One expert highlighted that he would also like to directly select suspicious cases in the Time Panel by selecting bars instead of using the temporal filter (see Figure 2.1, A.1). Another suggestion was to provide the possibility to put suspicious receivers on a black list to stop and investigate all transactions to these receivers. Furthermore, one expert suggested to have a simulation feature to being able to play around with the composition, weights, and thresholds of sub-scores and see how it affects the overall score in order to optimize the scoring algorithm.

However, all three investigators propose to include: support of network analysis. Network visualization is highly interesting in the area of fraud detection in order to analyze the relations between accounts, receivers, and dimensions. Such a network visualization could help them to see the connection between suspicious transactions and other transactions as well as involved accounts.

In addition to the mentioned opportunities for improvement we observed several minor usability issues (e.g., labels were sometimes too small or they overlapped) which will be resolved in the next iteration of EVA.

2.7 Discussion

In this section, first, we discuss which features presented in EVA fulfill each of the defined requirements. Next, we illustrate the challenges and opportunities that we identified during the work process.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

2.7.1 Requirements

The set of views presented in EVA supports a better understanding of the data by presenting it in different abstraction levels to the investigator. Identifying time-oriented aspects, analysing the score construction, as well as drill-down inspection are tasks that can be executed simultaneously in different views. EVA also presents fully responsive interaction techniques that allow a natural understanding of the relationships between the multiple coordinated views. Aiming to perform a more concise decision about an alarmed transaction, the investigator can explore data features, identify patterns, and evaluate scores by selecting and filtering the views. All interaction that excludes or include data into a view ensures the consistency of the data by also excluding, including, or rearranging data representations on the other views as well.

The time-oriented analysis is presented through the views A.1 and A.2 and through their link with the other views. Investigators can observe sending, receiving, scores, amount and others feature patterns over the time. Different periods of time can be specified during the analysis. The multivariate feature of transactions makes a complete visual encoding of all important characteristics within a single view impossible. By presenting multiple-coordinated views we present an overview of different features that our collaborators declared essential for their tasks. However, by presenting a dynamic table in addition to these multiple views, EVA allows detailed analysis of raw data features. Although EVA provides features to compare multiple customer accounts, it was primarily developed for the analysis of individual accounts. The simultaneous investigation of multiple accounts usually includes transactions of a time span from one to three years. However, due to national law restrictions, the maximum period of time that a bank is allowed to keep this transaction data is seven years. Thus, even for the most extreme outlier case present in our real world data (2,000 transactions per year), EVA scales fine.

In view B we can analyse how each overall score is constructed by others specific scores (**R1**). By observing how each line passes through the axis it is possible to identify which sub-scores influence most. In addition, once all transactions are visible at once in this view, EVA allows an analysis of the usual construction pattern for an account. Thus, when analyzing a single transaction, the investigator can compare the construction of it with the normal construction behaviour. Another feature is the range filters for any of the axis. By doing that, the investigator can judge if high/low sub-scores are being ignored, overestimated, or present correlation and, thus, help on evaluating the scoring system.

In order to aid multiple account comparison and analysis, view E present an interactive row chart that allows the investigators to filter each account‘s transactions (**R2**). By doing that, it is possible to compare patterns among the different views presented by the prototype.

Similarities and differences between suspicious and non-suspicious transactions become more evident during visual analysis. This can be done for single or multiple accounts. Thus, transactions that are wrongly flagged as suspicious (false positive cases) can be

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

more easily identified when comparing them with other transactions from the same account **(R3)**. On the other hand, false-negative cases are also more easily detected by the exploration of the suspicious patterns through visual means **(R4)**. A false-negative case is illustrated and described in Figure 2.4. An efficiency way to perform both tasks is by using the Dynamic Table (F) in combination with the Score Construction View (B).

2.7.2 Solving Real-World Tasks with EVA

We chose the tasks for our evaluation session in order to reflect the investigators' real world tasks (we elaborated them together with two collaborating domain experts). In this section, we outline how an investigator solved his real-world tasks with EVA and the insights he derived during the evaluation session. While one investigator used EVA to solve Task 2, he examined account acc10421 (see Figure 2.4). Interactively investigating the different views of EVA, he used the Scores Construction View to filter out transactions with a low suspiciousness of the country the money was transferred to (see arrow in Figure 2.4, view B). For this filter selection only one transaction was not automatically flagged as suspicious (i.e., the gray element in Figure 2.4) and EVA shows several clues that indicate that the only non-flagged transaction might be fraudulent too and should at least be considered as suspicious. All transactions (including the non-flagged transaction) are going out to the same receiver (see arrow in Figure 2.4, view D.1), on the same date (see arrow in Figure 2.4, view A.2). In addition, the non-flagged transaction involves a high amount of money (see arrow in Figure 2.4, view C) and scored quite high in several sub-domains (see Figure 2.4, view B). However, the overall score was not high enough to flag this transaction as suspicious. After some more exploration the investigator confirmed this transaction as a false-negative case. Without the VA support of EVA, it would have been (nearly) impossible to spot this mistake of the automatic scoring system, which illustrates the benefits of a VA approach compared to a pure FFD approach. This insight led the investigators to actually fine-tune the automatic alert system (see Section 2.4.4).

2.7.3 Limitations & Further Work

Demands to detect, analyze, and monitor suspicious behavior are constantly increasing not only in FFD. Based on some of EVA's limitations, we present possible further work and open research challenges.

Network Analysis. Although our prototype shows promising results for investigating long time intervals of transactions and relating a small number of accounts, we do not support the investigation of networks of accounts yet. An interactive network visualization would allow investigators to better reason about suspicious money transfer relationships and patterns within their contexts [Cha+07].

New Customer Classification. When a new customer is added to a profile system, he/she does not have enough transactions to derive a reliable profile by EVA's profile generation algorithm (see Section 2.4.4). This makes it impossible for the scoring algorithm to detect fraudulent attempts from new accounts. There is a need for an

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

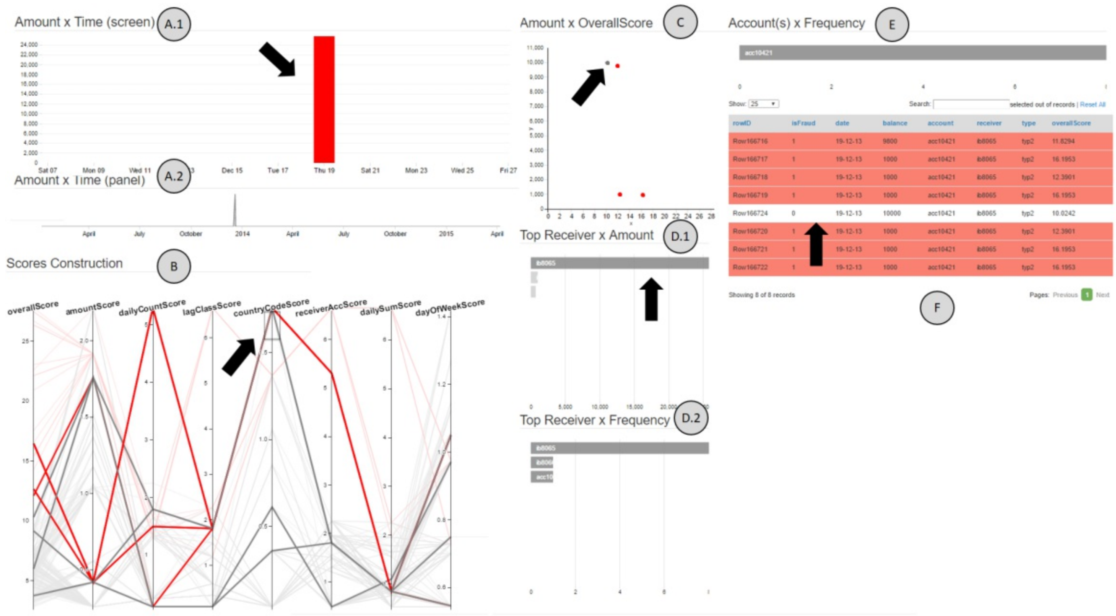

Figure 2.4: False-negative case spotted during the evaluation of EVA. We composed this figure to show different steps in the analysis process. Black arrows indicate interesting insights that are discussed in subsection 2.7.2.

exploratory VA environment which allows for the analysis of suspicious behavior of new customer accounts.

Knowledge Base Construction. One aspect that adds up to the complexity of fraud detection is that finding suitable solutions for detecting and deciding about suspicious cases is not enough [DR15]. Currently, investigators are using their experience to judge if a transaction is fraudulent or not. This can be tricky and results vary with investigators. We suggest the construction of a knowledge base based on former fraud detection that supports investigators to choose suitable scoring thresholds during analysis. For instance, by logging investigators' interaction data, we could collect filter set ups or filter combinations that obtained most success on detecting fraudulent behavior. This would not only support fine-tuning of the automatic scoring system but also ease the knowledge transfer to inexperienced investigators. Furthermore, it could also keep all investigators updated on new fraudulent discoveries.

Multiple Customers Monitoring. In our work, we can handle a small group of customers. Usually, these are accounts that were flagged as suspicious by the automatic scoring system. However, it should be possible to monitor all customers (or at least a big parts of them), which could lead to new insights. During evaluation (see Section 2.6.1) investigators also suggested to keep a "black list" of fraudulent customers to block them as well as to use them as blueprints for the identification of new fraudulent attempts.

Fine-Tuning Fraud Detection Algorithms. EVA's Score Construction view already

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

facilitates reasoning about how suspiciousness scores were constructed by sub-scores and which combinations of sub-scores are well suited as indicators. This feature also supports investigators in evaluating their algorithm (see Section 2.7.2). However, directly manipulating the weights of sub-scores for overall score construction is not supported by EVA at its current state. A visual interactive support for fine-tuning the algorithm would be a valuable addition to this work.

Different Types of Fraud. EVA was designed and implemented to meet the specific needs of the investigators at our collaborating financial institution, and thus, to tackle a specific kind of financial fraud, i.e., identifying and analyzing unauthorized transactions. VA support for detecting other types of fraud (see Section 2.4.3) is still needed, which prompts important research challenges for further work.

2.8 Conclusion

During the development of EVA we followed an iterative design over a period of 1.5 years in close collaboration with domain experts from a national bank. EVA follows the VA principles of interweaving intuitive interactive visualizations and analytical techniques in a seamless way. We selected our visualization techniques as well as the interaction techniques with special consideration of our design requirements, derived from discussions with our collaborating domain experts, who had limited experience with visual exploration tools. Therefore, we decided to use well-known interactive visualization techniques our experts are mostly familiar with. In the background we use the automatic computation of profile-based suspiciousness scores, which helps to monitor the behavior of costumers (in particular, payment transactions, which are characterized by multivariate and time-oriented aspects). We pursued an interactive multi-coordinated view approach, which took Tufte's [Tuf11] principles of graphical integrity into account. In particular, we obeyed Tufte's principle of *"show data variation, not design variation"* [Tuf11] (page 61). EVA eased the overall FFD process, which was enjoyable verbalized by one investigator during the evaluation sessions as *"...it is very interactive and you can browse the data, even if you don't know what you are looking for, and find new insights"*.

We evaluated EVA with real world data and could demonstrate that EVA was able to scale well even for extreme cases and to perform the required tasks in a suitable and appropriate way. The analysis of the insights, that were discovered the study participants, supported a more comprehensive understanding of the professional usage of EVA. Participants predominantly looked for connections in the data. They primarily got these connections from comparing different views, but also from comparing different variables. It was fascinating to observe that the number of connection insights decreased with the complexity of the task while other types of insights (coincidence, contradiction) emerged. Moreover, this real case of a overlooked fraudulent transaction was discovered during this evaluation study due to the indisputable benefits of visual exploration. We are aware that transferability from three study participants is limited, but nevertheless, this analysis indicates that future research in this area might lead to striking results

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

concerning insight generation with VA approaches in relation to the complexity of tasks.

Based on our study, we also propose possible future research directions in the field. Since the tasks involved in FFD are similar in different event detection domains, our approach may be transferable to other domains too, such as malware risk analysis, health parameter monitoring, terrorist detection, and governmental fraud.

2.9 Acknowledgements

This work was supported by Centre for Visual Analytics Science and Technology CVAST (funded by Austrian Federal Ministry of Science, Research, and Economy in the exceptional Laura Bassi Centres of Excellence initiative, project number: 822746).

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

Bibliography

[AA06] N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. ISBN: 978-3-540-31190-4. DOI: 10.1007/3-540-31190-4_3. URL: http://dx.doi.org/10.1007/3-540-31190-4_3.

[Aig+11] W. Aigner et al. *Visualization of time-oriented data*. Springer Science & Business Media, 2011.

[Car+14] M. Carminati et al. „BankSealer: an online banking fraud analysis and decision support system“. In: *IFIP International Information Security Conference*. Springer. 2014, pp. 380–394.

[Cha+07] R. Chang et al. „Wirevis: Visualization of categorical, time-varying data from financial transactions“. In: *Visual Analytics Science and Technology. VAST. IEEE Symposium on*. IEEE. 2007, pp. 155–162.

[Cle84] W. S. Cleveland. „Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging“. In: *The American Statistician* 38.4 (1984), pp. 270–280.

[DML14] M. Dumas, M. J. McGuffin, and V. L. Lemieux. „Financevis.net - A Visual Survey of Financial Data Visualizations“. In: *Poster Abstracts of IEEE VIS 2014*. Poster and Extended Abstract. Nov. 2014.

[DR15] W. N. Dilla and R. L. Raschke. „Data visualization for fraud detection: Practice implications and a call for future research“. In: *International Journal of Accounting Information Systems* 16 (2015), pp. 1–22.

[Ers] Erste Bank, Austria. *Erste Group IT International*.

[HB03] M. Harrower and C. A. Brewer. „ColorBrewer. org: an online tool for selecting colour schemes for maps“. In: *The Cartographic Journal* 40.1 (2003), pp. 27–37.

[HLN09] M. L. Huang, J. Liang, and Q. V. Nguyen. „A visualization approach for frauds detection in financial market“. In: *Information Visualisation, 13th International Conference*. IEEE. 2009, pp. 197–202.

[Ise+13] T. Isenberg et al. „A systematic review on the practice of evaluating visualization“. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2818–2827.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

[Jai11] K. Jaishankar. *Cyber criminology: exploring internet crimes and criminal behavior*. CRC Press, 2011.

[Kei+08] D. Keim et al. „Visual Analytics: Scope and Challenges". In: *Visual Data Mining*. Ed. by S. J. Simoff, M. H. Böhlen, and A. Mazeika. Springer, 2008, pp. 76–90. DOI: 10.1007/978-3-540-71080-6_6.

[Kir+99] J. D. Kirkland et al. „The NASD Regulation advanced-detection system (ADS)". In: *AI Magazine* 20.1 (1999), p. 55.

[Kle13] G. Klein. *Seeing what others don't: The remarkable ways we gain insights*. PublicAffairs, 2013.

[KMH06a] G. Klein, B. Moon, and R. R. Hoffman. „Making Sense of Sensemaking 1: Alternative Perspectives". In: *IEEE Intelligent Systems* 21.4 (July 2006), pp. 70–73.

[KMH06b] G. Klein, B. Moon, and R. R. Hoffman. „Making Sense of Sensemaking 2: A Macrocognitive Model". In: *IEEE Intelligent Systems* 21.5 (Sept. 2006), pp. 88–92.

[Ko+16] S. Ko et al. „A Survey on Visual Analysis Approaches for Financial Data". In: *Computer Graphics Forum*. Vol. 35. 3. Wiley Online Library. 2016, pp. 599–617.

[Kou+04] Y. Kou et al. „Survey of fraud detection techniques". In: *Networking, sensing and control, 2004 IEEE international conference on*. Vol. 2. 2004, pp. 749–754.

[KP15] S. Kriglstein and M. Pohl. „Choosing the Right Sample? Experiences of Selecting Participants for Visualization Evaluation". In: *EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (EuroRV3)*. Ed. by W. Aigner, P. Rosenthal, and C. Scheidegger. The Eurographics Association, 2015. DOI: 10.2312/eurorv3.20151146.

[KTM09] J. Kielman, J. Thomas, and R. May. „Foundations and frontiers in visual analytics". In: *Information Visualization* 8.4 (2009), p. 239.

[Lei+15] R. A. Leite et al. „Visual Analytics for Fraud Detection and Monitoring". In: *Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on*. IEEE. 2015, pp. 201–202.

[Lei+17] R. A. Leite et al. „EVA: Visual analytics to identify fraudulent events". In: *IEEE transactions on visualization and computer graphics* 24.1 (2017), pp. 330–339.

[Lue10] J. Luell. „Employee fraud detection under real world conditions". PhD thesis. University of Zurich, 2010.

[MA14] S. Miksch and W. Aigner. „A matter of time: Applying a data–users–tasks design triangle to visual analytics of time-oriented data". In: *Computers & Graphics* 38 (2014), pp. 286–290.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

[Mac86] J. Mackinlay. „Automating the design of graphical presentations of relational information". In: *ACM Transactions On Graphics (Tog)* 5.2 (1986), pp. 110–141.

[Mica] Microsoft. *Excel*. office.microsoft.com/en-us/excel/ (accessed: 2016-12-09).

[Micb] Microsoft. *PowerPoint*. office.microsoft.com/en-us/powerpoint/ (accessed: 2016-12-09).

[Mon+13] M. Monroe et al. „Temporal event sequence simplification". In: *IEEE transactions on visualization and computer graphics* 19.12 (2013), pp. 2227–2236.

[Mun09] T. Munzner. „A nested model for visualization design and validation". In: *IEEE transactions on visualization and computer graphics* 15.6 (2009), pp. 921–928.

[PC05] P. Pirolli and S. Card. „The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis". In: *Proceedings of International Conference on Intelligence Analysis*. 2005, pp. 2–4.

[PSM12] M. Pohl, M. Smuc, and E. Mayr. „The User Puzzle—Explaining the Interaction with Visual Analytics Systems". In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (Dec. 2012), pp. 2908–2916.

[Ric02] D. J. H. Richard J. Bolton. „Statistical Fraud Detection: A Review". In: *Statistical Science* 17.3 (2002), pp. 235–249. ISSN: 08834237.

[Rin+13] A. Rind et al. „Interactive information visualization to explore and query electronic health records". In: *Foundations and Trends® in Human–Computer Interaction* 5.3 (2013), pp. 207–298.

[SK95] J. P. Steidlmayer and G. Kummel. *Financial data event flow analysis system with study conductor display*. US Patent 5,454,104. Sept. 1995.

[Tuf11] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 2011.

[Wag+15] M. Wagner et al. „A Survey of Visualization Systems for Malware Analysis". In: *EG Conference on Visualization (EuroVis) - STARs*. Ed. by R. Borgo, F. Ganovelli, and I. Viola. The Eurographics Association, 2015, pp. 105–125. DOI: 10.2312/eurovisstar.20151114. URL: http://mc.fhstp.ac.at/sites/default/files/publications/wagner_2015_eurovis_star_malwarevis_postprint.pdf.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

CHAPTER 3

NEVA: Visual Analytics to Identify Fraudulent Networks

This paper first appeared in: Computer Graphics Forum, Volume 39, number 6 pp. 344–359, 2020. ©2020 Computer Graphics Forum. Reused with Permission (See [A L+20])

Full List of Authors Besides Roger A. Leite, the following authors contributed and provided valuable input to this paper: Theresia Gschwandtner, Silvia Miksch, Erich Gstrein, and Johannes Kuntner.

3.1 Abstract

Trust-ability, reputation, security, and quality are the main concerns for public and private financial institutions. To detect fraudulent behavior several techniques are applied pursuing different goals. For well-defined problems, analytical methods are applicable to examine the history of customer transactions. However, fraudulent behavior is constantly changing, which results in ill-defined problems. Furthermore, analyzing the behavior of individual customers is not sufficient to detect more complex structures such as networks of fraudulent actors. We propose NEVA, a Visual Analytics exploration environment to support the analysis of customer networks in order to reduce false-negative and false-positive alarms of frauds. Multiple coordinated views allow for exploring complex relations and dependencies of the data. A guidance-enriched component for network pattern generation, detection, and filtering support exploring and analyzing the relationships of nodes on different levels of complexity. In six expert interviews, we illustrate the applicability and usability of NEVA.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

3.2 Introduction

Detecting and understanding the network of related events is an important task in several domains such as biology, medicine, insurance companies, retail companies, public sector, and banks. Consider the following sample questions: "How are those diseases linked?", "What is the influence of this component on that treatment?", "Which selling products are linked, how, and why?" (see "the parable of the beer and diapers" case [Whi11]). In financial institutions, network analysis is used to understand users' behavior and to detect frauds. They often use automatic algorithms to detect fraudulent events of single actors, such as assessing if a transaction fits the transaction history of the respective customer. The goal of such a technique is to reduce the amount of false-positive and false-negative findings, to avoid harm in a magnitude of hundreds of thousands of dollars. However, this type of individual analysis system is sensitive and needs to be in constant evaluation in order to ensure the quality of the results. In this work, we focus on the detection and analysis of fraudulent networks of bank transaction events. We analyze visual patterns for different types of frauds (i.e., unauthorized transactions, money laundering, and straw persons). Besides dealing with a data type with complex features such as **time-oriented** and **multivariate** aspects [Aig+11], we propose the exploration and reasoning about a **network** of individuals.

Pattern and outlier detection are common tasks of Artificial Intelligence (AI) approaches in financial fraud detection (FFD). The main challenge of applying AI techniques to FFD is the constant adaptions required due to the creativity and fast change in fraudulent strategies which lead to false-negative findings. False-positive and false-negative findings come at expensive costs. While false-positive alarms might lead to the accusation of innocent people, false-negatives mean that fraudsters succeed to cause financial harm to the bank or to innocent bank customers. While these are two different circumstances, the fine-tuning of detection algorithms for both need to be done in combination. More sensitive algorithms lead to the detection of more frauds but also generate more false-positive alarms. On the other hand, a more generic algorithm would detect fewer frauds, and thus, more false-negative alarms would be generated. Consequently, the calibration of their sensitivity is essential. Thus, in order to improve FFD, it is important (1) to constantly adapt and balance the parameters of algorithms to identify as many fraudulent cases as possible without triggering too many false alarms and (2) to support informed reasoning about the output of these algorithms, such as potential false positive and false negative results.

Types of Financial fraud are classified with respect to different features, such as the amount of money, time of the transaction, and the relations of involved accounts. Besides applying AI for pattern and outlier detection, one of the main challenges on FFD is detecting fraudulent networks. However, performing sub-graph search is an extremely demanding task and it is characterized as an NP-complete problem. Even with punctual optimization available [Nei12; Spe+16], this task would require a sweep of all nodes and their relation layers recursively. Hence, this results in costly algorithms. Since the negative impact of fraudulent attacks increases over time, the task of FFD is of social

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

and financial importance and it must be detected as fast as possible.

Tasks of fraud detection are an open problem that require visual exploration, discovery, and analysis [KTM09]. VA would offer great benefits by integrating human analysis into this complex process [Kei+08]. However, the current solutions for FFD mainly use data mining techniques, with just a few exceptions involving VA techniques. Thus, we propose a VA approach for the investigation of fraudulent networks, based on an automatic FFD alert system. In this work, we focus on detecting specific types of financial frauds, such as “straw persons” and “money laundering” within a financial institution. We designed this approach with respect to Munzner’s nested model [Mun09], which makes it flexible and extensible enough to be adapted to similar domains with similar multivariate, relational, and time-oriented aspects. We also used the “Design Triangle” by Miksch and Aigner [MA14] to identify the target audience, define the data model, and find the important tasks in the target domain. Thus, our main contributions are:

- Designed in close collaboration with domain experts, NEVA improves the network analysis for FFD by intertwining automatic methods and visualization techniques within an interactive exploration environment (Section 4.5).
- To the best of our knowledge, we present a new guidance-enriched component for network pattern generation, detection, and filtering that supports different levels of analysis complexity (Section 3.5.4).
- We illustrate the applicability and usefulness of NEVA with four real-world tasks and discuss the lessons learned from an evaluation session with six domain experts (Section 3.6 and Section 4.6).
- We identify and elaborate on open challenges and possible future research directions in the field (Section 4.7).

3.3 Related Work

This work is mainly motivated by the same problem discussed in [**2018_leite**]. However, a network exploration and analysis perspective is brought by NEVA, to better contextualize that we categorize related works into the following seven topics:

VA for the financial domain. Looking for visual approaches for financial data, we identified FinanceVis [DML14] which is a survey that provides a browser tool that includes over 85 articles related to financial data visualization. Moreover, Ko et al. [Ko+16] presented a survey of approaches for exploring financial data. Motivated by a lack of information available, financial data experts were interviewed about their preferences regarding automated techniques, visualizations, data sources, and interaction methods. This survey highlights the many under-examined financial business domains and argues for the need of more works presenting design, development, and results involving real-world financial data, as our approach does.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Areas of Fraud detection. There are several state-of-the-art reports with emphasis on general fraud detection. One of the first modern analysis approaches concerning fraud detection was published by Bolton and Hand [Ric02] in 2002. They identified four areas: credit card fraud, money laundering, telecommunication fraud, and computer intrusion. The same types of fraud were described by Kou et al. [Kou+04] but are broadly classified into: misuse and anomaly detection. These works supported our understanding of the diverse fraud domains and their common approaches. However, no analyzed work presented a multivariate, time-oriented, network VA approach for FFD of banks transactions.

Financial fraud detection (FFD). One of the pioneers of designing visual solutions to support FFD analysis was Kirkland et al. [Kir+99]. They presented NASD, which uses visual techniques to facilitate the interpretation of detected frauds. NASD's Regulation Advanced-Detection System (ADS) is supported by five different visualizations. Moreover, ADS combines detection and discovery features that can support different domains. This work presents AI for pattern recognition, visualizations to aid human reasoning, and data mining to support regulatory analysis. While this was pioneer work in the field of FFD, they do not live up to state-of-the-art features such as interactive exploration of the visualized data. Argyriou et al. [ASV14] aimed to find fraudulent activities committed by employees of a company (internal frauds). In contrast to that, we aim to investigate other types of frauds: fraudulent schemes involving a network of users (i.e., money laundering schemes).

Another approach for FFD is presented by WireVis's [Cha+07], using multiple coordinated views. The main idea is to explore big amounts of transaction data through interactive views in order to aid fraud detection. The approach highlights similarities between accounts based on transaction keywords over time. WireVis clusters a set of accounts based on their similar keywords, depicting relationships among accounts and keywords over time. Thus, this approach is limited to analyzing transactions that present similar keywords in its description. Due to the similar data type and the use of multiple connected (and interactive) views, this is the most similar approach to NEVA. However, NEVA is especially focused on the detection and analysis of fraudulent networks, and thus, provides specialized means for network pattern analysis.

Our previous approach, *complementing work* [**2018_leite**], tackles the problem from a different direction and presents an integration of VA methods into an existing "detection and decision" workflow. This approach combines automatic methods with well-known visualization techniques in order to lower the learning effort for domain experts. We developed our approach following the same "familiar well-known visualization" design thinking. Different from EVA [**2018_leite**], NEVA aims FFD support to frauds investigations involving network aspects such as "money laundering" and "straw person". On the other hand, EVA [**2018_leite**]'s main goal was the discovery and reasoning about frauds coming from individuals history analysis as "unauthorized transactions".

Automatic methods for FFD. A first financial flow analysis approach is presented by [SK95]. This work focuses on data aggregation in order to allow users to draw analytical

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

conclusions and make stock decisions. A more modern approach, EventFlow [Mon+13] is a query and data transformation tool for temporal event data sets designed to facilitate analysis. This approach provides aggregated data visualization representations to track events that are related over time.

Recent needs in FFD are presented by Dilla et al. [DR15]. The authors presented theoretical framework to predict how the investigators should apply VA techniques. They evaluated various visualization techniques according to different cognitive processes. The discussion about the benefits of interactive data visualization for fraud detection was one of the main discussion topics, which was also used as one of the main points of our research. A decision support based on profile generation and analysis is presented for online banking fraud analysis from Carminati et al. [Car+14]. This semi-supervised approach provides no visual support for fraud analysis. However, it is directly related to our approach since we are also focusing on profile analysis for fraud detection. This approach has a strong statistical meaning. We believe that VA methods have great potential to improve the investigation of the FFD and enable the analyst to better understand the lacks of the scoring systems.

Networks in finance. Franklin Allen and Ana Babus [AB09] presented a non-visual paper that discusses a node-link representation as being a natural financial system representation which can efficiently explain certain economic phenomena. Tekušová et al [TK08] presents a node-link diagram that is generated using economic analysis methods. The system aims at showing patterns for large corporate shareholder networks. It allows the visual analysis of cash flows and the identification of shareholders. When targeting event monitoring works, Huang et al. [HLN09] presented a VA framework for stock market security. Aiming for a reduction of false alarms produced by traditional AI techniques, this work presents a visualization approach combining a node-link diagram for network analysis and a 3D treemap for market performance analysis. Didimo et al. [DLM14] presents a VA tool that allows the analysis of different institutions and also the analysis of internal transactions of a bank is considered. The paper describes the two different types of network to correctly model the data.

One of the biggest influence to develop our approach comes from the work of Cheng et al. [Che+17]. This VA approach to loan network risk management presents risk measurements by analyzing subgraphs flows in a bigger network. In this work, 20 subgraphs are analyzed during the study case. Even subgraph search being an NP-complete problem there are algorithms that try to optimize the performance of the task [Spe+16; ZC16]. These subgraphs models coupled with the subgraph search challenge inspired us to create a "Node-Link Patterns" concept that allows us a subgraph exploration with guidance (see Section 4.5). Wang et al. [Wan+18] proposed a system of multiple coordinated views that support anomaly detection for global trade network analysis. The system takes localization and events (i.e., armed conflicts) into consideration and relates them to international trades.

Graph query. Some works have been demonstrating the efficiency of using visual language to filter graphs in order to find expected patterns and results [Pie+17; Pie+16].

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

On the other hand, VIGOR [Pie+18] is focused on supporting users to better reason about graph query results (grouping nodes, adding labels to clusters, and so on). Parts of our approach allow also a "drawable" query design similar to [KZA10; Cha+08]. TeFNet aims to contrast tax evasion, money laundering, and fiscal frauds [Did+19]. The query language and the visualization techniques rely on a suitable timeline approach that maps time to space. The approach presents efficient results for large graphs. Like NEVA, the system has been tested in a real working environment. However, NEVA provides means for "drawable" queries (see Section 3.5.4) developed especially for FFD. Moreover, different from the auto-completion guidance during query writing implemented by VISAGE [Pie+16], NEVA adds guidance support during pattern drawing/querying (see Section 3.5.4). After discussing with FFD experts, we decided to exclude approximate results as GRAPHITEs [KZA10] does only precise results in order to not mislead any search or pattern understanding.

Guidance started in the area of Human Computer Interaction (HCI) [Hor99] and was recently characterized in the context of VA by Ceneda et al. [Cen+17]. Ceneda et al. define: "Guidance is a computer-assisted process that aims to actively resolve a knowledge gap encountered by users during an interactive visual analytics session". The authors extend van Wijk's model of visualization, and thus, they present a general model to facilitate in-depth reasoning about guidance. To illustrate their model with examples, they use existing guidance approaches from the literature. Three distinct degrees of guidance are described by Ceneda et al.: (1) Orienting, techniques that support building and preserving a mental map, (2) Directing, presents to the user a pre-selection pool of possibilities (a recommendation system is an example), and (3) Prescribing, techniques that force the user to take the recommended next step. In a literature review about guidance in visual data analysis [CGM19], the authors highlighted the different guidance degrees (Prescribing, Directing, and Orienting) which can be used to classify guidance. The guidance-enriched features implemented in NEVA, present a degree of Directing guidance because they provide a set of alternative options that guide the analyst to avoid the formulation of useless queries (see Section 3.5.4).

3.4 Financial Fraud Detection (FFD)

We designed, developed, and evaluated our approach in collaboration with a national bank [Ers]. Our main goal was to improve the current FFD techniques used by our partner institution. However, before we present NEVA (Network dEtection with Visual Analytics), we briefly discuss (1) the characteristics of transaction data, (2) the complexity of detecting networks of fraudsters, (3) a summary of the currently used pipeline for FFD, and (4) the scoring approach used to identify suspicious transactions. Since we give a detailed description of these aspects in one of our previous works [**2018_leite**], we just summarize the main aspects here in order to ground our contribution.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

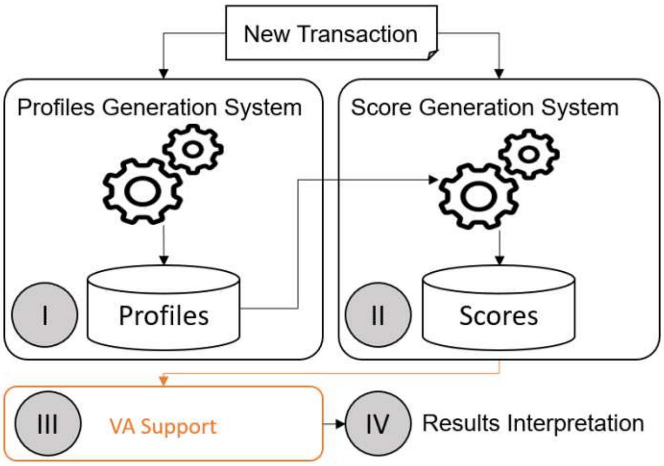

Figure 3.1: The design overview of the transaction evaluation system. We present a new interactive VA approach to support the investigation of suspicious behavior. We highlight in orange how our approach fits into the workflow of FFD (III).

3.4.1 Transaction Data

We used real data of money transactions from our bank partner for the development of our approach. The data from a two years period that represents payments and money transfers were anonymized and internally checked by the bank before they made it available to us. The transaction events have several categorical, numerical, geospatial, and temporal dimensions (i.e., money sender/receiver, amount of money, location, time of execution, and others). While our approach is focused on investigating relational features, we also support the exploration of the most critical data dimensions (such as the amount of money) that were identified by our collaborating domain experts. Moreover, we defined different categories of node types such as money senders and receivers based on Cheng et al. [Che+17] (see Section 3.5.4).

3.4.2 Problem Complexity

Fraudsters tend to adapt their fraudulent strategies continuously. This is why automatic algorithms, for instance searching for a pre-defined set of patterns, is unlikely to succeed in the context of FFD. Automatic means can be used to identify abnormalities, however, a human investigation is required to confirm the harmfulness of suspicious cases. To support this task, VA techniques enable the human to interact with the data and to improve the reasoning process. In addition to frequent pattern changes of fraudulent attempts, several other aspects add to the task complexity of FFD:

In one of our previous works [**2018_leite**], where we approach a related problem of FFD and investigate data with the same characteristics, we go into details about: (1) the **scalability** complexity of dealing with hundreds of thousands of transactions per day, (2) the **context** complexity of understanding the motivation behind a financial crime,

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

(3) the complexity of **time-oriented data analysis** [Aig+11] that might obfuscate some frauds, and (4) the problem of **fraud classification** which might include a huge number of sub-classes of frauds due to the multivariate nature of the data. All these aspects were addressed during NEVA's design and implementation.

3.4.3 Methodology for FFD

In this subsection, we give an overview of the workflow pursued by our collaborators to identify and reason about financial frauds. According to the privacy policy of our collaborating bank, we cannot go into details about the actual fraud detection algorithm. However, we roughly sketch the four steps of the used methodology: profile generation, score generation, results interpretation, and fraud validation.

Profile Generation. For each customer, an automatic system generates a profile based on the transaction history of his/her account (see Figure 3.1 I). Profile generation is a process that has its own rules of execution and it is not synchronized with the other steps of the workflow (i.e., profiles are re-computed every week or every 10 transactions).

Score Generation. Each incoming transaction is compared with the respective customer's profile (see Figure 3.1 II). For a new transaction the algorithm considers data dimensions such as: operation location, operation time, amount of money, and other features to compare the new action with the usual behaviour of that customer (i.e., the customer's profile). Then, the algorithm generate sub-scores for each dimension and summarize them into one overall score. The higher the score, the more suspicious is the transaction.

Results Interpretation. This is the non-automatic phase of the investigation where the investigators analyze multiple transactions simultaneously, due to time constraints. Transactions whose scores are over a given threshold are further filtered by predefined rules. Based on their personal experience, the investigators then decide whether an alarm should be considered fraudulent or not (see Figure 3.1 IV).

Fraud Validation. After deciding for a suspicious transaction, a further personal investigation is required. The bank might stop the transaction in some cases. From this step on, there are different legal approaches that might be applied according to the type of fraud that it seems to be.

3.4.4 Types of Frauds

While types of frauds are manifold, in this subsection, we briefly describe the types of frauds that we focus on.

Money Laundering aims to transform money from crime and corruption into legal money. Mostly, this type of fraud involves a network of accounts. **Straw Persons** main goal is using someone with low suspiciousness levels to get illegal money for someone else who is not legally allowed to receive it. **Unauthorized Transactions** involves

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

transactions that were sent from a customer account but were not authorized by the account owner.

3.4.5 Scoring Approach

Due to security and privacy constraints of our collaboration partners from the bank, we cannot describe the scoring algorithm in detail. However, we compare it to another public approach from Carminati et al. [Car+14]. In this approach, customer profiles are generated in a semi-supervised way outputting several kinds of statistical measurements. This approach flags transactions as suspicious. When evaluating our scoring algorithm by comparing it to the old scoring approach used by the bank, we could confirm that domain experts could detect 500% more confirmed fraudulent transactions, preventing 86% of the financial losses [**2018_leite**].

3.5 Design and Implementation

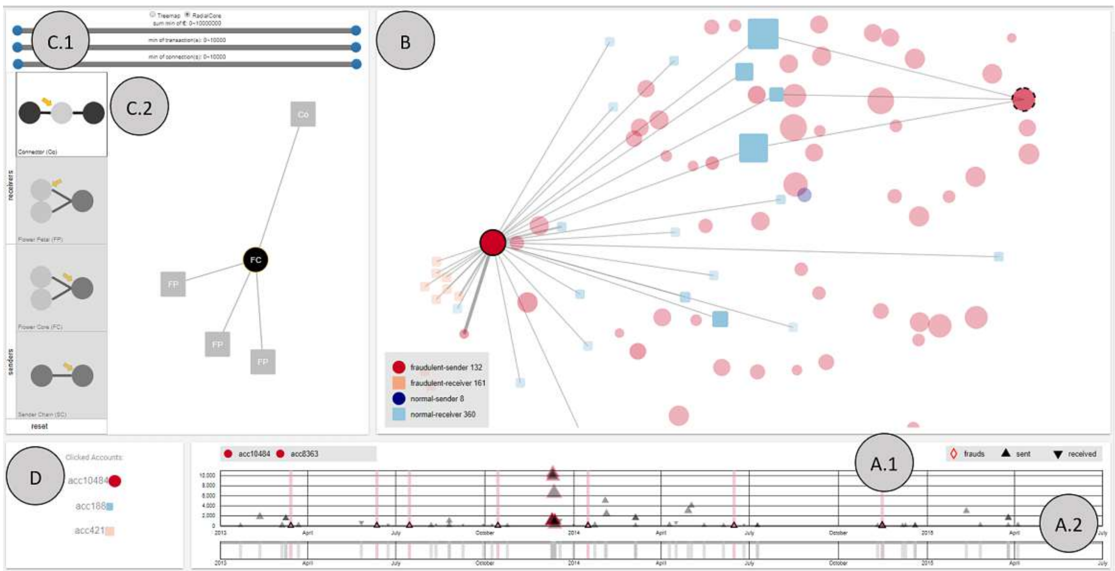

Figure 3.2: Screenshot of NEVA (Network dEtection with Visual Analytics). (A.1, A.2) Temporal Views: the two views present temporal detail (A.1) and overview (A.2) information. (B) Node-Link View: this view represents the network of the analyzed bank accounts. (C.1, C.2) Guidance-Enriched Pattern Search: In this panel, we allow filtering the accounts with respect to the connections between them by using sliders (C.1) and changing the layout of the network through radio buttons (C.1). Moreover, we provide a guidance-enriched network pattern specification, detection, and filtering approach (C.2). (D) This view presents the history of inspected nodes for keeping track. In all views, elements that represent suspicious data are highlighted in red.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek, Your knowledge hub, WIEN

We design NEVA by using a combination of two methodologies: "Nested Model" [Mun09], and "Design Triangle" [MA14]. While the nested model methodology guarantees a flexible and extensible interactive approach, the design triangle methodology supports a better understanding of the scope of the problem by the clarification of the Data, User, and Tasks.

Data. Financial transaction events constitute multivariate, relational, and time-oriented data, including details about the transactions such as amount, time, sender, receiver, etc.

Users. Investigators from financial institutions that investigate and validate alerts of suspicious transactions.

Tasks. The overall tasks are network related fraud detection and reasoning by means of profile analysis. These tasks include mainly the reduction of false-negative and false-positive alarms, history comparison, as well as the manual investigation of suspicious transactions.

3.5.1 Requirements

In the context of fraud detection and the analysis of financial networks, there are still many unresolved challenges [Kou+04; AB09]. With our VA solution, we focus on supporting the tasks of "results interpretation" and "fraud validation" (see Section 3.4.3). To support fraud investigators during a better decision-making process, in collaboration with the domain experts, we agreed to address the following requirements:

R1: Identification of False-Negative Alarms. Aiming to overcome the lack of accuracy of automatic algorithms, the exploration and identification of hidden fraudulent accounts is an important task that should be supported by the proposed solution. The system needs to facilitate the reasoning about potential criminal networks. It should not only be possible to investigate suspicious transactions but also the accounts linked to them. By visualizing non-flagged accounts (i.e., accounts that were not flagged as suspicious by the automatic detection mechanism) in relation to flagged accounts (i.e., possible fraudulent accounts), investigators might be pointed to suspicious behavior that could not be detected by automatic algorithms.

R2: Identification of False-Positive Alarms. False-positives alarms are transactions that were flagged by the automatic system as potential frauds but actually they are not. The amount of false-positive alarms varies according to the calibration of the automatic system. However, 100% precision cannot be achieved (i.e., a good trade-off has to be found between the number of false-positive and false-negative alarms), and thus, there will always be false-positives. Flagged transactions, however, lead to the investigation of the involved accounts and their related network. This group of accounts need to be inspected in order to reason about the potential fraud, which leads to a further decrease of false-positive alarms. To support informed decision making during the validation step (see Section 3.4.3), the system has to support the visual analysis of the network of flagged accounts and its related network.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

R3: Identification of Different Types of Frauds. With the current system, it is possible to automatically identify unusual behavior. However, the classification of this behavior is still an open issue. Classification, however, is essential to better understand the possible consequences of an alarm and make better-informed decisions on how to handle it. The system should support the investigation of suspicious behavior of accounts that are involved in flagged transactions and their related network. Thus, based on the money flow of accounts, the identification of different types of frauds should be possible.

R4: Guided Network Pattern Exploration and Search. The changing behavior of fraudulent attempts needs to be constantly monitored. Thus, our system needs to provide efficient support to reason about identified patterns of possible fraudulent attempts. Moreover, searching for specific types of frauds (i.e., patterns) within the comprehensive network of financial transactions should be supported. However, searching for arbitrary patterns within a huge network of financial transactions might easily lead to cognitive overload. Thus, the user needs guidance on which patterns are worth investigating.

3.5.2 Data Setup

Aiming to improve the current methodology for FFD (see Section 3.4.3), we kept our main focus on analyzing suspicious transactions. Thus, instead of displaying the 77,000 accounts at once, we propose an automatic initial filter based on the investigators' workflow. Thus, we suggest a step-wise approach to identify additional fraudulent cases that could not be identified by automatic means. First, (1) we apply a filter on the data set to select just accounts that contain alarming transactions. Next, (2) we select accounts that had relations to at least two suspicious accounts. This process is called "man-in-the-middle" selection (see an example of a "man-in-the-middle" in Figure 3.8). Lastly, (3) we analyze the relations of all found "man-in-the-middle" accounts that are not only receptors but also senders of money in order to find newly hidden "man-in-the-middle" accounts (this case is represented in Figure 3.7).

While the majority of accounts selected in this phase might not seem suspicious at first sight, the investigators consider that "man-in-the-middle" accounts represent a certain logistic risk and require further investigation. Those accounts are likely to be part of fraudulent schemes. This analysis was previously conducted by analyzing spreadsheets. So, the data presentation made by our VA approach in addition to the automatic FFD mechanism enhances the possible analysis scope by allowing investigators to interact visually with the data.

3.5.3 Network Event Detection with Visual Analytics (NEVA)

The main idea of our proposed approach started as a further work proposed by [**2018_leite**]. As a sequel to our former work, we again developed this approach following the iterative design process [Mun09]. We had several meetings with our collaborators to discuss the data, the context, the problem, and the tasks. During the design process, we created

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

a number of visual encoding designs and asked for expert feedback. We iteratively designed several prototypes with which we could test these designs, discard ideas, and refine interactions. Our approach was developed using web technology (HTML, CSS, Javascript), as well as services and libraries such as JQuery, D3.js, and Google Material.

NEVA is composed of four main coordinated views that display different aspects of the data (see Figure 4.3). All views are connected interactively. We opted for well-known visualizations in order to keep the learning curve as low as possible, and thus, to foster acceptance. However, all visualization techniques were chosen with respect to their suitability for the data and the tasks at hand [Mac86]. Next, we discuss the motivation of the different visual design choices made in our approach referring to A.1, A.2, B, C.1, C.2, and D from Figure 4.3.

A.1, A.2: Temporal Views. The combination of both views (see Figure 3.3) supports a **detail** and **overview** visual analysis of time-oriented aspects of the data [CKB09]. With the main focus of representing an overview, the view A.2 is smaller and supports brushing and filtering, which are then reflected in the detail view (A.1). By brushing A.2, the user zoom in the corresponding time gap in A.1 and, thus, potential over-plots in the time representation (A.1) are spaced and clarified. In both views, rectangles highlight points in time (represented by the x-axis) at which a transaction happened. These rectangles present a color opacity feature that when accumulated over each other is decreased and, thus, facilitates the observance of over-plots. The two views combination allows for analyzing temporal details while preserving context information.

We opted for a 1D representation in the **overview** view A.2 in order to keep it simple. This view serves two main functions, (1) giving an overview of all events of the data set and (2) allowing to zoom in and zoom out of the detail view A.1 with a brush interaction. Due to that, we constructed this view with just enough space on the y-axis for brushing and filtering at an overview level.

The **detail** view A.1 was designed to allow for analyzing temporal aspects of transactions, represented by triangles. While the orientation of the triangles represents if a transaction was sent (triangle up) or received (triangle down), the amount of money sent is represented by its position on the y-axis. The triangle's size encodes the suspiciousness score of the transaction and a red stroke indicates if a transaction's score is above a given threshold. Hovering the triangles provides detailed information about the transactions. For representing multiple transactions, we opted to present a glyph-based approach in this view to better represent the collection of sent and received transactions. We sort all transactions before plotting them, from highest to lowest amount of money. Thus, small transactions, which result in small glyphs, are always plotted in the last layer. We tried bar charts but, when zooming out, one bar would need to aggregate multiple transactions (losing information, as line charts would do too) or to be very thin (losing visual appeal and adding difficulty to interactivity).

B: Node-Link View. We use visual variables to encode several features in the node-link diagram. **Link width** encodes the number of transactions between two nodes. **Node**

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

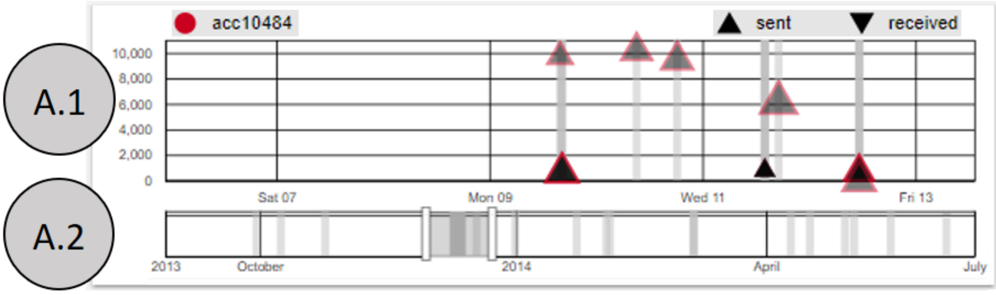

Figure 3.3: The two temporal views A.1 (detail) and A.2 (overview). The selected account is displayed at the top left corner, while the glyph legends are displayed on the top right corner. The investigators can also interact with the legend to filter the view.

size encodes how many connections (i.e., transactions from or to other accounts) a certain node has. In our scenario the accounts that are investigated by means of this node-link visualization are filtered to those accounts that have been flagged as suspicious and other accounts that are connected to these suspicious accounts. Thus, the more connections a node has within this subset of possibly fraudulent accounts, the more suspicious it is itself. Thus, we use size to make them visible and to highlight big players. We used **node color** to represent four categories of accounts: orange nodes represent suspicious receiver accounts, red nodes represent suspicious sender accounts, light blue nodes are receivers that are non-suspicious so far, and dark blue nodes are non-suspicious senders. Since one of the main interest of the investigators is to distinguish (1) "senders" accounts from (2) "receivers" accounts, to assure a good differentiation of sender and receiver nodes, which are already differentiated by dark and light color shades, we decided for double encoding these also to **node shape**: circles (senders) and squares (receivers). **Node transparency** represents how many suspicious transactions an account has compared to its total amount of transactions. The more suspicious transactions are associated with an account, the less transparent we draw the corresponding node. **Node stroke** is used during interactions to highlight selected or hovered nodes. Interaction-wise, the investigator can select nodes, drag the camera view, and zoom in and out of the view. When a node is selected, the view shows all first layer nodes with links and all second layer nodes without its links. Selecting second layer nodes for comparison is also supported.

We included two different force directed **layouts**: a treemap layout and a radial layout with cores (see Figure 3.4). A treemap-like subdivision of the space with four divisions that represent the amount of each account type (fraudulent receiver, fraudulent sender, normal receiver, and normal sender) and pulls these node types towards the center of the respective subdivision. While creating attraction centers at each region, the nodes also get influenced by node-link spring forces. The influence of the spring forces avoid that the nodes merge and clutter to the same region center points, even when they are being pulled in the same direction. The radial layout with cores assumes a light gravitational force in the middle in order to keep the elements close but the main aspect of it is the repelling force that the nodes apply to each other. By coupling the repelling force from

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

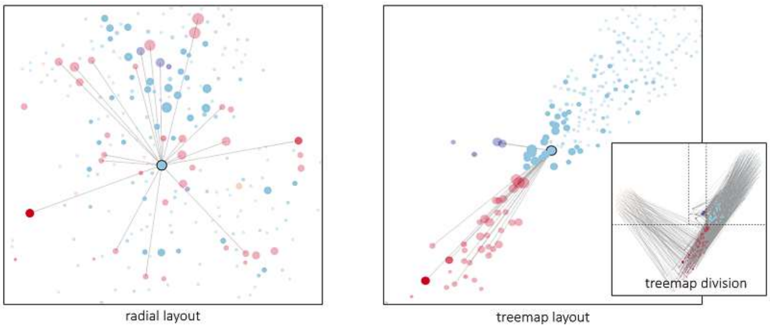

Figure 3.4: Two layout options presented by NEVA. On the left-hand side, a radial layout which is better suitable for path-finding tasks due to to the distribution of nodes on the available space. On the right-hand side, a treemap layout which groups the different types of accounts (fraudulent-senders, fraudulent-receivers, normal-sender, normal-receiver) into an underlying treemap segmentation of the available space.

all nodes with the link attraction force, we could obtain a network representation that visually separates sub-graphs of related events. Both layouts pre-calculate the position of nodes and do not allow drag and drop interactions in order to preserve the position of nodes, and thus, the mental map of the user.

To represent the network connections, we decided for a node-link view over a matrix view, since a node-link presentation better supports path-finding [OJK18]. This feature facilitates the reasoning about circular schemes such as money laundering. While the position of node elements for different layouts (which all highlight different aspects of the network) presents an additional challenge, node-link diagrams are better suited for reading information from small and sparse graphs [KEC06]. In addition, a matrix view would introduce challenges of sorting and positioning accounts so that relevant accounts are close to each other to enable comparison, which usually demands several sorts of interaction features. Another beneficial feature of node-link diagrams is that they enhance reasoning about indirect paths between two and more nodes [KEC06]. Moreover, a node-link representation provides several visual attributes that can be used to support effective visual analysis without resulting in a confusing visualization (i.e., link width, node size, node color, node transparency, or node stroke). Furthermore, we provide a sub-graph "drawing" feature (i.e., drawing node-link connections) for querying the data set (see D: Guidance-Enriched Pattern Search) which is a more intuitive approach than asking the investigator to draw a sub-graph matrix.

C.1, C.2: Filters and Guidance-Enriched Pattern Search. In this panel we provide filtering relations between accounts with respect to different features by using sliders (C.1), changing the graph layout through radio boxes (C.1), and a guidance-enriched pattern generation, detection, and filtering canvas (C.2). Based on the most common queries manually executed by investigators, we designed three sliders that can be combined to narrow down the content of an investigation (see Figure 3.5 C.1). These

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

sliders are used to filter network features (i.e., frequency of transactions and amount of money involved) and to specify filter intervals for these network features (min/max). Despite C.1 presenting a simple but very useful filter functionality, C.2 presents a more complex approach that is explained in Section 3.5.4.

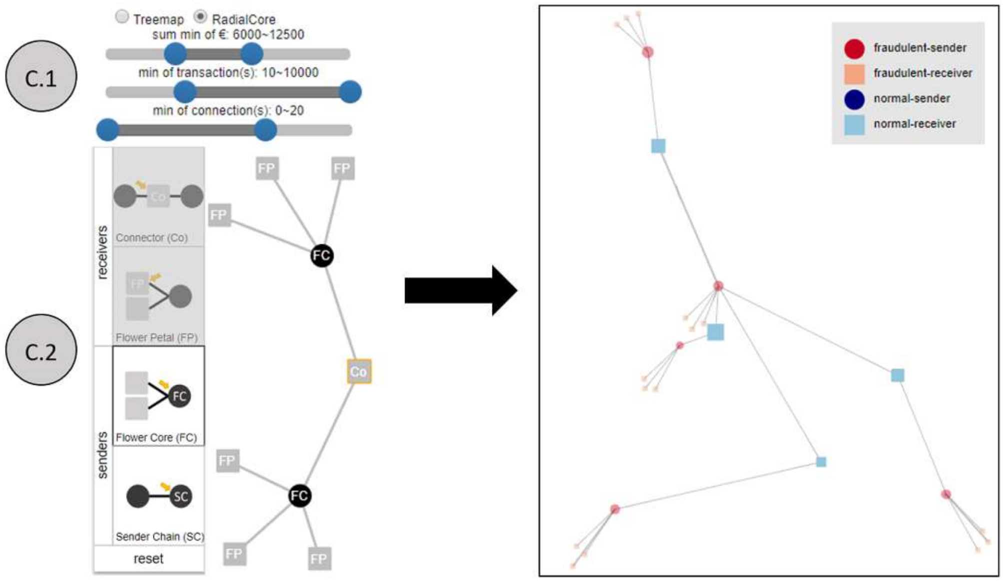

Figure 3.5: C.2 shows four possible sub-network structures to be selected. **Connectors (Co)** are receiver nodes that are in between sender nodes. **Sender Chains (SC)** are sender nodes that are connected to each other. **Flower Core (FC)** is a sender node which is just connected to receiver nodes. **Flower Petal (FP)** are receiver nodes that are only connected to one sender node. Overall, this is an example of possible patterns that can be drawn and queried. C.1 provides sliders to apply interval filters. In C.2, the investigator drew a network pattern to query for: the black nodes present two FC connected to three FP each. The investigator then connected the two FC by a Co node. At the right-hand side, we see the query result of the applied filters and the drawn pattern. Light blue nodes are **Co**s, dark blue nodes are **SC**s, red nodes are **FC**s, and orange nodes are **FP**s.

D: Selection History. This view helps to keep track of the network investigations that have been performed. Through this view, we also can re-select already investigated nodes.

3.5.4 Node-Link Patterns & Guidance-Enriched Pattern Canvas

Our approach to provide guidance for the detection of interesting patterns within the network was motivated by the work presented by Cheng et al. (see Figure 15 in [Che+17]). In their work, the authors elaborate that different combinations of only four nodes

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

generate nearly 200 different types of sub-graphs, however, just 20 of these sub-graphs could be found in the data set. Thus, as the number of different possible combinations of n nodes increases exponentially, we guide users in constructing sub-graphs that are used to query the data set. Showing them which combinations of patterns make sense because they are actually present in the data considerably reduces the number of possible sub-graphs.

Besides simplifying pattern searches by allowing the user to simply draw the sub-graph that should be queried, we also facilitate these sub-graph query constructions with guidance. Guidance is seen as a computer-assisted process that gradually narrows the knowledge gap that hinders effective continuation of the data exploration and analysis. It provides prospective assistance so that users can make sense of the data on their own [CGM19; Cen+17; Cen+18]. In analogy with the auto-complete text guidance assistant in VISAGE [Pie+16], we propose a solution that preserves the investigator's mental map. Common text auto-complete features might disrupt an investigator's mental map [PHG06; APP11] by suggesting well known or most used queries. Instead of influencing the query with our guidance system, we apply guidance only to avoid that queries will produce empty search results. Thus, we step-wise guide the user through all possibilities of sub-graph constructions that actually exist in the customer graph. Node combinations that do not exist in the data, cannot be drawn; only the ones that exist are available. Moreover, our visual query approach is more user friendly than a text-based approach that might require scripting knowledge. Based on Ceneda et al.'s definition of different guidance degrees [Cen+17; Cen+18], we categorize the guidance degree provided in view C.2 as: "directing guidance", since it narrows down the multitude of options which in theory are possible at each step of constructing a sub-graph, to those options that actually make sense. The input is network data as well as the user's actions (i.e., which node is currently selected) and the output are possible additions to the drawn sub-graph, which is used to query the data. In order to perform a guidance-enriched network pattern exploration and search function (see Section 3.5.1), first, we defined four different categories of nodes according to their relations (i.e., "Connector (Co)", "Flower Petal (FP)", "Flower Core (FC)", and "Sender Chain (SC)"). All four categories are described in Figure 3.5. Reducing a pattern representation to these categories of linked nodes proved to be very powerful by not only being able to represent all patterns described in Cheng et al.'s approach [Che+17] but also for being able to construct any money flow sub-graph existing in our real world data set. The complete representation is guaranteed for our use case because each node from our data set falls into one of the four classes. Thus, we use these node categories as a base for building the pattern search feature in view C.2. When it comes to performance, all sub-graph queries applied to our real-world dataset (see Section 3.5.2) during development and evaluation of NEVA, presented "instant" results (i.e., execution average of 3 ms when using a computer with Intel Core i5-2520 2.50 GHz with 8GB RAM).

In view C.2 the investigator can draw any sub-graph by adding and linking the previously proposed node-link patterns. When drawing such a sub-graph, NEVA guides

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

the investigator by providing possible links to add to the sub-graph, while avoiding non-existing constructs or links that make no sense. The options are automatically calculated based on the node that the investigator selects in the query canvas, and thus, the provided guidance is an active answer on the user's action and his/her knowledge gap (i.e., complete awareness of which patterns exist in the customer graph). In this way, our guidance support facilitates pattern generation, detection, and filtering tasks [Cen+17; Cen+18]. Therefore, we name it Guidance-Enriched Pattern Canvas.

Our guidance approach prevents two different types of common errors when constructing sub-graphs: First, if a structure cannot be found in the current data set, it is not possible to draw this structure in the Guidance-Enriched Pattern Canvas. For instance, let's assume that the maximum connections of any Flower Core node (FC) in a data set are three Flower Petal nodes (FP). After drawing the third link from a Flower Core node (FC) to a Flower Petal node (FP) in the drawing area of view C.2, the Flower Petal node (FP) selection option would be grayed out. The second type of error prevented by our guidance system is the combination of patterns and nodes that do not make sense (i.e., the logical structure and connection of different types of nodes). For instance, two Flower Petal nodes (FP) cannot be connected to each other, otherwise it would represent a Sender Chain node (SC). On the other hand, at least two Sender Chain nodes (SC) must be connected to each other. For each added node in the drawing area, the query is immediately executed and all other views are updated accordingly. With that in mind, we illustrate a real-world example in Figure 3.5. C.2 shows that a Connector node (Co) is selected (see the yellow bordered squared node) to extend the query pattern. Since a Connector node (Co) cannot be linked to another Connector node (Co) and neither to a Flower Petal node (FP), these options are grayed out.

Sub-graph search algorithm. We first categorize all nodes in the dataset by categories (FP, FC, SC, or Co). Once the nodes are categorized, we can search for matching subgraphs in the data model. We define $G_t = (V_t, E_t)$ as our data model graph, being V_t the vertex set and E_t the set of ordered pairs of nodes representing the graph edges. We also define $|V_t| = n$. The sub-graph pattern that the user constructs during a query is defined as $G_p = (V_p, E_p)$. Next, we use our pattern matching algorithm to look for a subgraph $G_r = (V_r \subseteq V_t, E_r \subseteq E_t)$. Considering that the G_p nodes represent different categories, which identify a set of nodes in V_t, this algorithm conducts the following operations: (1) for every node category $v_p \in V_p$ it extracts the nodes $v_t \in V_t$ that match the constraints defined in v_p and assigns them to V_r; (2) for every node now in V_r, it checks whether its edges connect to E_r. An edge $e_t = (u_t, v_t) \in E_t$ belongs to E_r if $u_t, v_t \in V_r$ and they were extracted from the two different categories nodes in V_p connected by an edge $e_p \in E_p$.

The algorithm has a quadratic asymptotic time complexity $O(n^2)$. The first step (1) has to be repeated for each node in V_p, therefore it has a complexity of $O(|V_p|)$. We can perform the second step (2) in $O(d)$ time for each node, being d the out-degree of a node. Therefore, considering that from V_p we extract the nodes to insert in V_r, the time complexity of the algorithm is $O(|V_r|d)$: $|V_r|$ is at most n (when the queried subgraph includes all the nodes in V_t) and d is at most $n - 1$ (when v_r shares an edge with all the

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

other vertices in the graph).

However, in practice, we could find that also with very dense graphs and real-world data our implementation was able to find subgraphs with an efficient time, enabling a real time interactive exploration. Even if the general pattern matching problem is NP-Hard, in our case the problem is greatly simplified by the pre-calculated categories that maps the nodes in the target graph.

3.6 Solving Real-World Tasks with NEVA

In this section, we aim to present four real-world insights about false-negative and false-positive cases that were **confirmed by investigators** using our approach.

3.6.1 False-Negative Alarms

When inspecting the "acc9711", one of the first network features that appeared to the investigator is its connection with another sender account, which would present a SC (see Figure 3.5). After a quick inspection of the linked fraudulent account "acc8315", the node-link view reveals that both accounts have the non-flagged account "acc9540" as a common receiver (see Figure 3.6). By observing the multiple transactions involving a high amount of money from the two connected fraudulent accounts, we conclude that the non-flagged account "acc9540" should also be considered suspicious in case of a fraudulent network scheme.

Selecting the "man-in-the-middle" (i.e., the non-flagged sender account "acc76") revealed that it received money from three different fraudulent senders (see red nodes in Figure 3.7). Since this account has suspicious connections and is also a sender account, the "man-in-the-middle" search algorithm was also applied to its connections. This revealed seven additional non-flagged senders that have common receptors with flagged accounts (see arrows in Figure 3.7). Further investigation confirmed the involvement of those accounts in a larger fraudulent network scheme. In this case, the "man-in-the-middle" selection highlighted a potentially fraudulent network of accounts that were not flagged as suspicious by the automatic detection mechanism.

3.6.2 False-Positive Alarms

The account "acc4238" presents two flagged money transactions on the same day to two different accounts. First, a 6,500 € transaction to account "ib3614" that resulted in a score of 25, and second, a 50 € transaction to account "ib3613" that resulted in a score of nine. While the first transaction is worry-some, especially due to the high amount of money involved, the second transaction involves just a small amount of money. Although the second transaction resulted in a much lower score than the first one, it still exceeded the threshold of the automatic detection algorithm. This sensitive behavior might be a result of the first alarm (from the same day) and can be neglected. Thus, this case

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

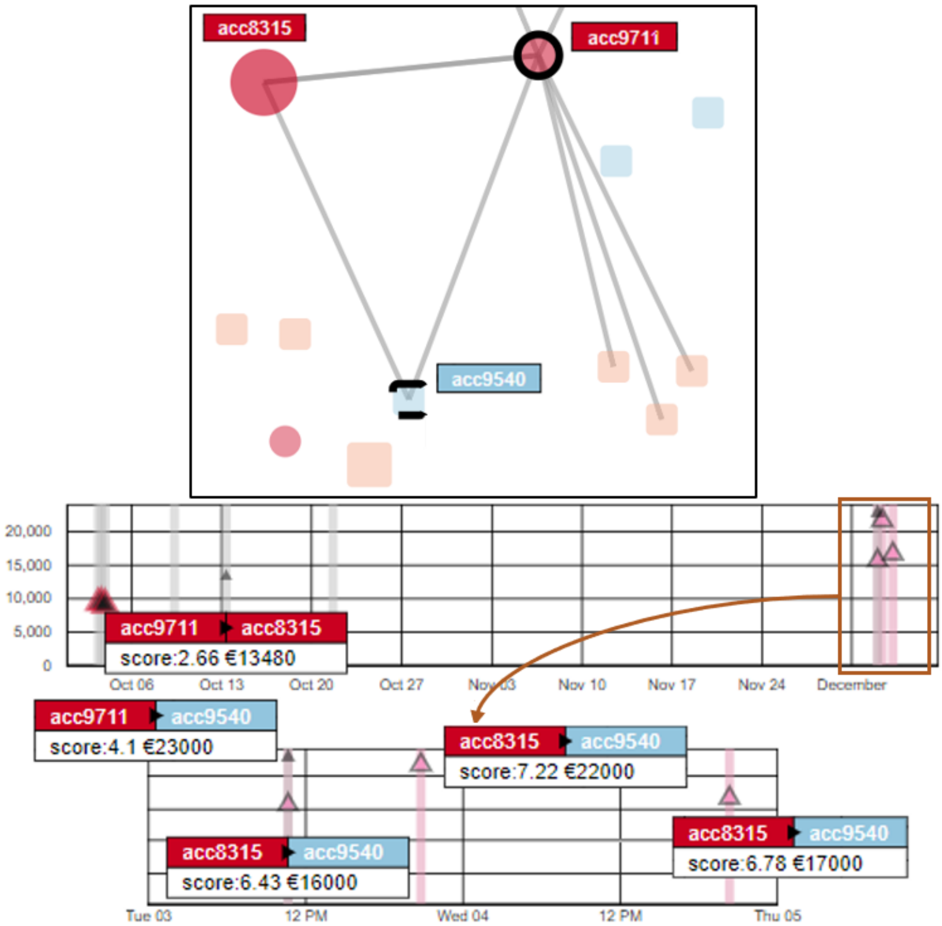

Figure 3.6: The triangle relationship between three accounts, i.e., the flagged accounts acc8315 and acc9711 as well as acc9540, an initially non-flagged account. It is possible to observe high-value transactions between the two flagged accounts and in addition, from both accounts to the same target account (acc9540). This indicates that acc9540 might be a fraudulent "straw person".

presents a false-positive alarm that would not be easily detected and justified without data exploration motivated by a VA tool.

While "man-in-the-middle" nodes are not flagged by the automatic detection mechanism, investigators still consider them as highly suspicious and used NEVA to further invest them individually. If an investigator analyses such a suspicious case and comes to the conclusion that it is harmless, it can be considered a false-positive alarm. In Figure 3.8 we present a case with a non-flagged node ("ib11196") being the "man-in-the-middle" of two different fraudulent accounts ("acc11598" and "acc30692"). The two transactions involved show a small amount of money (30 € each). Although frequently sending small amounts of money could indicate a certain type of fraud, NEVA shows that those two

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

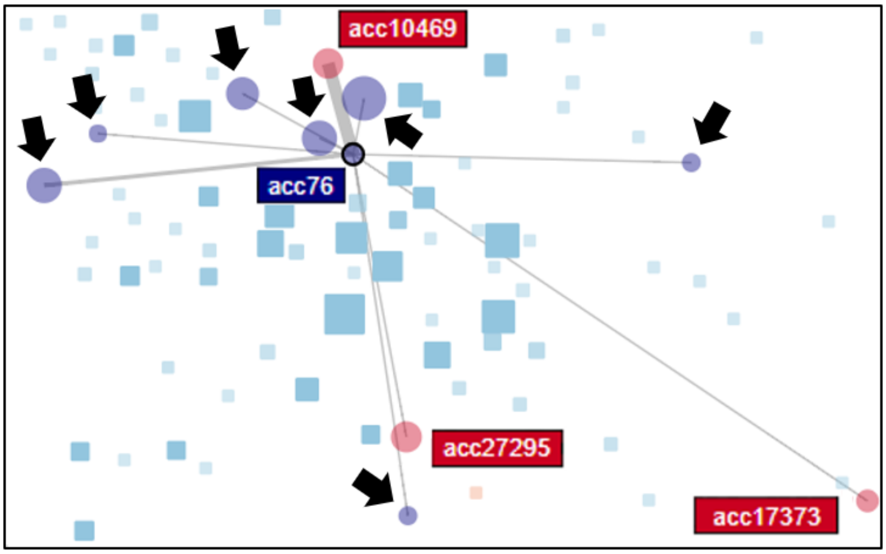

Figure 3.7: A case that acc76, a non-flagged sender, is detected as a “man-in-the-middle” node that connects three flagged senders. Moreover, it sent money to seven additional non-flagged accounts. Thus, NEVA automatically includes these non-flagged accounts for a comprehensive investigation of possible fraudulent networks (see Section 3.5.2).

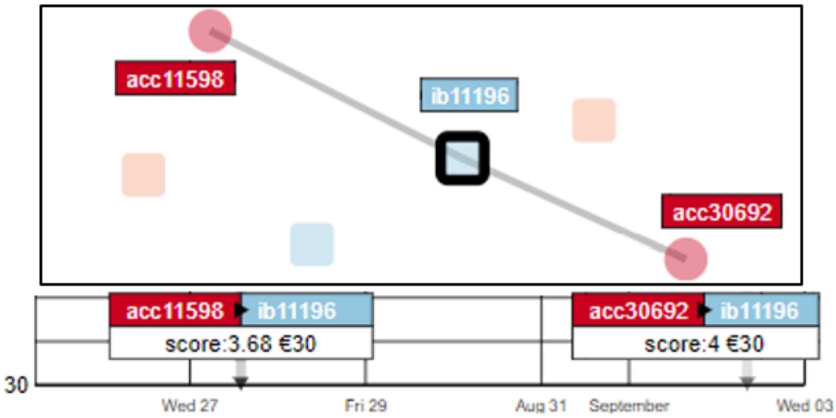

Figure 3.8: The selection analysis of the “man-in-the-middle” account “ib11196”. Although this account is connected to two fraudulent senders, the temporal view reveals a non-suspicious behaviour of only two received transactions of 30 €, which makes it very unlikely that this account is involved in any fraud.

relations do not have any previous or subsequent involvement. Thus, it does not present a critical account.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

3.7 Evaluation

In order to estimate the usability of our approach, we conducted a qualitative evaluation with six domain experts. By asking them to answer three research questions (RQ), a qualitative study allowed us to get the investigators feedback and reasoning about the insights that were gained while using NEVA.

RQ1: Comparison. What are the advantages and disadvantages of NEVA compared to the tools which investigators usually use?

RQ2: Insights. What kind of insights can be generated with NEVA?

RQ3: Improvements. Do the investigator miss any features or have suggestions for improvement?

Participants, Dataset, and Tasks. We recruited six potential investigators from our collaborating bank who had never seen the prototype before. Qualitative studies are a useful means to generate insights with relatively few study participants [Ise+13; KP15]. All participants had previous experience with visualizations for data presentation.

During the evaluation, we used an anonymized real-world **dataset** from our collaborators covering an interval from January 2013 to April 2015. After the data setup (see Setion 3.5.2), 661 different accounts with a total of 1,527,583 transactions of different types (i.e., net banking transactions and automatic payments of bills) were selected to be inspected. These tasks are structured according to the analytical task taxonomy by Andrienko and Andrienko [AA06], distinguishing between elementary and synoptic tasks.

To evaluate the **task** performance of our approach, we stipulate four tasks directly related to the proposed requirements (see Section 3.5.1). Task 1 is designed to evaluate R1, Task 2 is designed to evaluate R2 and so on.

Task 1: Identify false-negative alarms for any period of time.

Task 2: Identify false-positive alarms for any period of time.

Task 3: Find at least two types of potential frauds.

Task 4: Query any fraudulent pattern known to you.

Our evaluation design covers the highlighted problem complexities (see numerical references of Section 3.4.2): Task 3 and Task 4 are addressing **context (2)** and **fraud classification (4)** challenges, all tasks are addressing **time-oriented data analysis (3)**, and **scalability (1)** is put to the test by the usage of a real-world dataset with 77,000 accounts (see Section 3.5.2).

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Procedure and Collected Data Analysis. The study **procedure** took place in two locations, a meeting room at the university and a meeting room at the bank headquarters. The participant and the evaluation moderator were present. Additionally, audio and video recording software were used to support further data analysis. First, a short introduction to the study's goal and the meeting schedule were presented. Next, a semi-structured interview took place in order to better understand the current approach for network FFD and the participant's VA background. Then, the participant was invited to perform the proposed tasks using NEVA. While interacting with the prototype, the participant was encouraged to think aloud (RQ2). After this phase, another semi-structured interview was conducted in order to collect feedback (RQ1, RQ3).

After the evaluation session, we **analyzed the collected qualitative data** (notes, audio, and video recordings). In order to understand how effective our approach was supporting exploration and sensemaking (RQ2), we opted to use Klein's model (see also [KMH06a; KMH06b]) to interpret the collected data due to its broad [PSM12] intelligence analysis approach. For this reason, we adopted three categories from Klein [Kle13] to categorize notes, audio, and video recordings.

Connection. These are the resulted insights of the combination of two or more views. For example, two views showing different aspects of the same data elements.

Coincidence. These insights are results of events that in a first moment do not present obvious connections but through the viewers' eyes seem to have a relationship. For example, two data elements with a similar outlier position might have the same non-identified source.

Curiosity. These are motivation insights. For example, seeing one data element positioned in a different place when compared to the others, arouses the viewers interest in finding an explanation to that.

3.7.1 Results

All the participants were able to achieve satisfactory results on the four tasks with an average duration of 12 min. The first interview session (before trying the prototype) took about 15 min and the second interview (after trying the prototype) took about 25 min.

In this subsection, we present the results from the evaluation session with respect to our research questions.

RQ1: Comparison

The participants usually use visualization approaches for data analysis and presentation purposes. Some of the tools cited by the participants were: Microsoft Excel [Mica], Microsoft PowerPoint [Micb], and Tableau []. From the available visualizations of the mentioned tools they are most familiar with line charts, pie charts, and bar charts.

When comparing our approach to the other tools, some participants highlighted that interactivity of our approach was helping a lot with the filtering of "interesting elements"

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

and the analysis of relations. Concerning the system precision on representing and filtering, one participant commented: "just by looking at the graphs and interacting with it, I can detect a small number of suspicious nodes and gaining an initial understanding of them". Another participant commented: "The interactive technique is not only good for detecting suspicious patterns but also to exclude false alarms". Besides that, many aspects of our approach such as the design choices of the node-link diagram and the temporal view, the history track panel, the slider filters, and mainly the network pattern drawing tool were very well received. The participants also demonstrated an interest in showing our solution to business, meaning that a potential official tool could be created based on our prototype. With that and other motivating statements, we reason that NEVA is a positive improvement to the current workflow of FFD investigators.

RQ2: Insights

Three kinds of insights were counted during the evaluation session: Connection, Coincidence, and Curiosity. We present the distribution of a total of 178 identified insights (average of 29 for each participant) according to the performed tasks in Figure 3.9. **Connection**. The dominant appearance (47.75%) of this insight is mainly due to the multiple coordinated views of our approach which link the node-link view (B), the temporal views (A.1, A.2), and the Guidance-Enriched Pattern Search (C.1, C.2). For example, observing a node in the node-link view would lead the investigator to select it and analyze its transactions in the temporal view. Another example is that investigators constantly checked changes in the node-link view when drawing a pattern with the network pattern drawing tool. **Coincidence**. With the smallest appearance (14.60%), this insight predominantly occurred during Task 3. Since this task was about investigating different types of frauds, it is plausible that this insight type is about finding hidden relations, played a main role. **Curiosity**. During the investigation of false-negative cases (Task 1), this was the main insight type with six appearances. By interpreting suspicious transactions (i.e., links) in the node-link view, the investigators often inspected also the nodes to check if the transactions were really suspicious. This represented (37.64%) of the total insights.

	Connection	Coincidence	Curiosity	= Total
Task 1	28	7	24	**59**
Task 2	13	5	17	**35**
Task 3	30	12	22	**64**
Task 4	14	2	4	**20**
= Total	**85**	**26**	**67**	**178**

Figure 3.9: Insight summary. The appearance and the sum of each insight type with respect to the tasks performed.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek, WIEN Your knowledge hub

RQ3: Improvements

During the last part of the evaluation sessions, the participants were encouraged to make suggestions about how the approach could be improved from their point of view. The first suggested feature was to allow excluding nodes and transactions in case the investigator finds false-positive alarms. Another feature suggested was an “undo” button that would return the whole system to how it was one action before. A mouse-over window for hovering edges, giving an overview of the frequency of transactions between two nodes before inspecting it in more detail in the temporal views was also desired. Concerning the temporal views, a participant recommended to add a selection box that would allow the investigator to observe only the interaction between the two selected nodes instead of only highlighting them visually. We consider all other suggested improvements as minor usability issues that will be added in the next iteration of our approach.

3.8 Discussion

In this section we discuss (1) how our approach fulfills the previously defined requirements, (2) benefits of a potential integration with EVA [**2018_leite**], and (3) limitations of our approach as well as future research challenges in this field.

3.8.1 Requirements

We designed and developed NEVA to support a guided exploration and informed reasoning about fraudulent networks in the field of FFD. Our solution provides different levels of data abstraction. Moreover, it supports important tasks in the context of FFD, such as (1) inspecting relations between bank accounts, (2) analyzing temporal aspects of financial transactions with detail and overview information, (3) filtering the network of accounts with respect to the characteristics of relations between accounts, (4) querying for specific network patterns supported by novel guidance mechanisms, and (5) identifying straw persons and ”humans-in-between” by ranking all accounts connected to a selected account in two layers: direct connections (1st level connections) and all connections of these first level connections (second layer connections). Moreover, all views are connected, i.e., they consistently reflect all changes applied to the data (i.e., filtering, selections, modifications).

Investigators can analyze temporal aspects using an overview and detail approach of the temporal views A.1 and A.2 (Figure 4.3). Since it is very difficult to encode all relevant data aspects in one view, these temporal views allow for the detailed inspection of transactions while preserving the bigger picture. Another feature that supports reasoning about the temporal aspects of transactions is the filtering consistency between the views. The combination of these views helps to reason about the temporal sequence and consistency of sent and received transactions of a selected account. Moreover, our approach also enables direct comparison of the transaction histories of two accounts. This

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

feature supports the investigator in reasoning about a potential fraudulent collaboration between two or more accounts.

Since we used a real-world data set for development and evaluation, our approach scales to the scope of data sets typically used for real-world FFD tasks. Due to local law restrictions, bank institutions are allowed to keep records for a maximum period of seven years. Thus, our approach scales well even to the highest possible number of transactions existing in available data sets (32,175 transactions).

The node-link view allows for an analysis of connections between flagged and non-flagged accounts. After identifying suspicious relations, by using the temporal view, we can reveal transactions with suspicious features and/or to suspicious accounts. This enables investigators to investigate these accounts and make better informed decisions about the suspiciousness of these accounts, and thus, to reveal potentially false-negative cases **R1**. Similar inspections can be performed the other way around, starting from accounts flagged as suspicious and investigating the reasons for which this account was flagged by the automatic detection mechanism. For example, during the evaluation session (Task 2), one investigator filtered the data displayed to a low amount of money transactions (using view D) and quickly inspected the remaining accounts. After a few minutes, s/he could already confirm that a good part of those accounts presented false-positives alarms **R2**.

The identification of different types of fraudulent behavior **R3** is a very difficult and sensitive task that requires the combination of all views of NEVA. The differences of account relations become more evident during visual analysis. In addition, the understanding and identification of fraud patterns might be registered for further usage during a network pattern query **R4**. We also presented a tool (view D) to guide the search for relational patterns that might support investigators in formalizing queries and also in the categorization of fraud types.

3.8.2 Potential Integration with EVA [2018_leite]

Since we already designed and developed a FFD tool focused on different techniques, we strongly believe that NEVA would lend itself to be integrated with this approach. NEVA adds new features for network analysis and tools to the investigation process. Using interactions such as node selection, the investigator could link the two approaches in order to perform different analysis without losing the investigation track. Thus, as an extension, the best of both approaches would sum up in a better and more complete solution.

3.8.3 Limitations & Further Work

Network pattern analysis and outlier detection are constant research challenges. Additional layers of complexity should be considered when involving time-oriented and multivariate data. Based on current limitations we derive research challenges to inspire future works.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

New Accounts Monitoring. Since our automatic algorithm is based on scores which are calculated based on the account's history, we cannot efficiently evaluate newcomer accounts. Moreover, it is known that some frauds involve completely new accounts in order to hide from these algorithms. However, for those frauds other algorithms apply. That being said, we strongly believe that FFD could benefit from solutions for (1) analyzing first steps of new accounts individually and (2) keeping track of accounts that cannot be interpreted by a history-based system. Another possible challenging future work would be (3) to support the migration of new accounts analysis for a score-based approach.

Subgraph Search Limitations. The Guidance-Enriched Pattern Search (view C.2) provides a more intuitive search and better understanding of the data through its interactive and responsive exploration. However, this (i.e., the subgraph isomorphism problem) is an NP-complete problem. For the analysis of larger data sets, the usability of this technique depends on the available computation power and/or advances in the state of the art of sub-graph matching algorithms. However, as we did in NEVA, the performance of the query algorithm can be improved when the attributes of nodes are used to constrain searches.

Temporal Analysis of Multiple Accounts. NEVA's temporal views (A.1 and A.2) already support the temporal analysis and inspection of transactions. With these views, we provide support for comparing two accounts. In future work it would be interesting to include a temporal view that supports the analysis of all accounts of the selected network in a non-cluttered way.

Collaborative Investigation. Some analysis scenarios, such as a governmental investigations, can be at huge scale and involve many investigators over a grand period of time. Many countries recently faced cases of political corruption involving illegal money transactions that take years of analysis. Other examples of huge investigations are Wiki Leaks [BHM13] and analysis of cryptocurrency networks. These types of investigations might demand for a collaborative analysis. Distributing workload and double checking hypotheses would potentially present faster and more objective results.

3.9 Conclusion

Based on our experience in this field and in tight collaboration with domain experts from a national bank, we iteratively designed our VA solution for FFD network analysis, called NEVA. Our approach follows the VA principles of intuitive and interactive visualizations in combination with analytical techniques. All design and interaction choices were made with special consideration of the previously defined requirements and with respect to the limited experience of FFD investigators with visual exploration tools. Our approach consists of an automatic computation of profile-based suspiciousness-scores for each transaction in combination with and interactive multiple-coordinated view approach to explore and reason about the behavior of bank accounts. By showing "data

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

variation" [Rob07; WWK00] NEVA improves the quality and speed of the investigation process.

We designed and evaluated NEVA with real-world data and six real-world domain experts, which helped to assess the required scalability of the approach. The added value of our approach is supported by (1) the analysis of insights gained with the help of NEVA during our evaluation session with real-world investigators (see Section 4.6) and (2) the additional fraudulent cases identified with the help of NEVA that were later confirmed by investigators to constitute actual cases that demand for further investigation (see Section 3.6). The results of the evaluation session helped us to assess the positive impact of our approach in a real-world setting. The improvement of task performance and the number of insights gained while using NEVA confirmed its usability and efficiency.

Based on our study results, we also propose possible future research directions that not only would be of added value to the FFD domain, but also to fields that share similar problems and data characteristics (i.e., multivariate, time-oriented, and network data aspects), such as biology, medicine, insurance companies, or the public sector.

3.10 Acknowledgments

This work was supported by Centre for Visual Analytics Science and Technology CVAST (funded by Austrian Federal Ministry of Science, Research, and Economy in the exceptional Laura Bassi Centres of Excellence initiative, project number: 822746).

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Bibliography

[] *Tableau*. https://www.tableau.com/ (accessed: 2018-10-09).

[A L+20] R. A. Leite et al. „NEVA: Visual analytics to identify fraudulent networks". In: *Computer Graphics Forum*. Vol. 39. 6. Wiley Online Library. 2020, pp. 344–359.

[AA06] N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. ISBN: 978-3-540-31190-4. DOI: 10.1007/3-540-31190-4_3. URL: http://dx.doi.org/10.1007/3-540-31190-4_3.

[AB09] F. Allen and A. Babus. „Networks in finance". In: *The network challenge: strategy, profit, and risk in an interlinked world* 367 (2009).

[Aig+11] W. Aigner et al. *Visualization of time-oriented data*. Springer Science & Business Media, 2011.

[APP11] D. Archambault, H. Purchase, and B. Pinaud. „Animation, small multiples, and the effect of mental map preservation in dynamic graphs". In: *IEEE Transactions on Visualization and Computer Graphics* 17.4 (2011), pp. 539–552.

[ASV14] E. N. Argyriou, A. Symvonis, and V. Vassiliou. „A fraud detection visualization system utilizing radial drawings and heat-maps". In: *Information Visualization Theory and Applications (IVAPP), 2014 International Conference on*. IEEE. 2014, pp. 153–160.

[BHM13] B. Brevini, A. Hintz, and P. McCurdy. *Beyond WikiLeaks: implications for the future of communications, journalism and society*. Springer, 2013.

[Car+14] M. Carminati et al. „BankSealer: an online banking fraud analysis and decision support system". In: *IFIP International Information Security Conference*. Springer. 2014, pp. 380–394.

[Cen+17] D. Ceneda et al. „Characterizing guidance in visual analytics". In: *IEEE Transactions on Visualization & Computer Graphics* 1 (2017), pp. 111–120.

[Cen+18] D. Ceneda et al. „Guidance or No Guidance? A Decision Tree Can Help". In: *EuroVis Workshop on Visual Analytics (EuroVA). The Eurographics Association*. 2018.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

[CGM19] D. Ceneda, T. Gschwandtner, and S. Miksch. „A review of guidance approaches in visual data analysis: A multifocal perspective“. In: *Computer Graphics Forum* 38.3 (2019), pp. 861–879.

[Cha+07] R. Chang et al. „Wirevis: Visualization of categorical, time-varying data from financial transactions“. In: *Visual Analytics Science and Technology. VAST. IEEE Symposium on*. IEEE. 2007, pp. 155–162.

[Cha+08] D. H. Chau et al. „Graphite: A visual query system for large graphs“. In: *Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on*. IEEE. 2008, pp. 963–966.

[Che+17] D. Cheng et al. „Visual analytics for loan guarantee network risk management“. In: *arXiv preprint arXiv:1705.02937* (2017).

[CKB09] A. Cockburn, A. Karlson, and B. B. Bederson. „A review of overview+ detail, zooming, and focus+ context interfaces“. In: *ACM Computing Surveys (CSUR)* 41.1 (2009), p. 2.

[Did+19] W. Didimo et al. „Visual querying and analysis of temporal fiscal networks“. In: *Information Sciences* 505 (2019), pp. 406–421.

[DLM14] W. Didimo, G. Liotta, and F. Montecchiani. „Network visualization for financial crime detection“. In: *Journal of Visual Languages & Computing* 25.4 (2014), pp. 433–451.

[DML14] M. Dumas, M. J. McGuffin, and V. L. Lemieux. „Financevis.net - A Visual Survey of Financial Data Visualizations“. In: *Poster Abstracts of IEEE VIS 2014*. Poster and Extended Abstract. Nov. 2014.

[DR15] W. N. Dilla and R. L. Raschke. „Data visualization for fraud detection: Practice implications and a call for future research“. In: *International Journal of Accounting Information Systems* 16 (2015), pp. 1–22.

[Ers] Erste Bank, Austria. *Erste Group IT International*.

[HLN09] M. L. Huang, J. Liang, and Q. V. Nguyen. „A visualization approach for frauds detection in financial market“. In: *Information Visualisation, 13th International Conference*. IEEE. 2009, pp. 197–202.

[Hor99] E. Horvitz. „Principles of mixed-initiative user interfaces“. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM. 1999, pp. 159–166.

[Ise+13] T. Isenberg et al. „A systematic review on the practice of evaluating visualization“. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2818–2827.

[KEC06] R. Keller, C. M. Eckert, and P. J. Clarkson. „Matrices or node-link diagrams: which visual representation is better for visualising connectivity models?“ In: *Information Visualization* 5.1 (2006), pp. 62–76.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

[Kei+08] D. Keim et al. „Visual Analytics: Scope and Challenges“. In: *Visual Data Mining*. Ed. by S. J. Simoff, M. H. Böhlen, and A. Mazeika. Springer, 2008, pp. 76–90. DOI: 10.1007/978-3-540-71080-6_6.

[Kir+99] J. D. Kirkland et al. „The NASD Regulation advanced-detection system (ADS)“. In: *AI Magazine* 20.1 (1999), p. 55.

[Kle13] G. Klein. *Seeing what others don't: The remarkable ways we gain insights*. PublicAffairs, 2013.

[KMH06a] G. Klein, B. Moon, and R. R. Hoffman. „Making Sense of Sensemaking 1: Alternative Perspectives“. In: *IEEE Intelligent Systems* 21.4 (July 2006), pp. 70–73.

[KMH06b] G. Klein, B. Moon, and R. R. Hoffman. „Making Sense of Sensemaking 2: A Macrocognitive Model“. In: *IEEE Intelligent Systems* 21.5 (Sept. 2006), pp. 88–92.

[Ko+16] S. Ko et al. „A Survey on Visual Analysis Approaches for Financial Data“. In: *Computer Graphics Forum*. Vol. 35. 3. Wiley Online Library. 2016, pp. 599–617.

[Kou+04] Y. Kou et al. „Survey of fraud detection techniques“. In: *Networking, sensing and control, 2004 IEEE international conference on*. Vol. 2. 2004, pp. 749–754.

[KP15] S. Kriglstein and M. Pohl. „Choosing the Right Sample? Experiences of Selecting Participants for Visualization Evaluation“. In: *EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (EuroRV3)*. Ed. by W. Aigner, P. Rosenthal, and C. Scheidegger. The Eurographics Association, 2015. DOI: 10.2312/eurorv3.20151146.

[KTM09] J. Kielman, J. Thomas, and R. May. „Foundations and frontiers in visual analytics“. In: *Information Visualization* 8.4 (2009), p. 239.

[KZA10] P.-Y. Koenig, F. Zaidi, and D. Archambault. „Interactive searching and visualization of patterns in attributed graphs“. In: *Proceedings of Graphics Interface 2010*. Canadian Information Processing Society. 2010, pp. 113–120.

[MA14] S. Miksch and W. Aigner. „A matter of time: Applying a data–users–tasks design triangle to visual analytics of time-oriented data“. In: *Computers & Graphics* 38 (2014), pp. 286–290.

[Mac86] J. Mackinlay. „Automating the design of graphical presentations of relational information“. In: *ACM Transactions On Graphics (Tog)* 5.2 (1986), pp. 110–141.

[Mica] Microsoft. *Excel*. office.microsoft.com/en-us/excel/ (accessed: 2016-12-09).

[Micb] Microsoft. *PowerPoint*. office.microsoft.com/en-us/powerpoint/ (accessed: 2016-12-09).

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

[Mon+13] M. Monroe et al. „Temporal event sequence simplification". In: *IEEE transactions on visualization and computer graphics* 19.12 (2013), pp. 2227–2236.

[Mun09] T. Munzner. „A nested model for visualization design and validation". In: *IEEE transactions on visualization and computer graphics* 15.6 (2009), pp. 921–928.

[Nei12] D. B. Neill. „Fast subset scan for spatial pattern detection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.2 (2012), pp. 337–360.

[OJK18] M. Okoe, R. Jianu, and S. G. Kobourov. „Node-link or Adjacency Matrices: Old Question, New Insights". In: *IEEE transactions on visualization and computer graphics* (2018).

[PHG06] H. C. Purchase, E. Hoggan, and C. Görg. „How important is the "mental map"?–an empirical investigation of a dynamic graph layout algorithm". In: *International Symposium on Graph Drawing*. Springer. 2006, pp. 184–195.

[Pie+16] R. Pienta et al. „Visage: Interactive visual graph querying". In: *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM. 2016, pp. 272–279.

[Pie+17] R. Pienta et al. „Visual graph query construction and refinement". In: *Proceedings of the 2017 ACM International Conference on Management of Data*. ACM. 2017, pp. 1587–1590.

[Pie+18] R. Pienta et al. „VIGOR: interactive visual exploration of graph query results". In: *IEEE transactions on visualization and computer graphics* 24.1 (2018), pp. 215–225.

[PSM12] M. Pohl, M. Smuc, and E. Mayr. „The User Puzzle—Explaining the Interaction with Visual Analytics Systems". In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (Dec. 2012), pp. 2908–2916.

[Ric02] D. J. H. Richard J. Bolton. „Statistical Fraud Detection: A Review". In: *Statistical Science* 17.3 (2002), pp. 235–249. ISSN: 08834237.

[Rob07] J. C. Roberts. „State of the art: Coordinated & multiple views in exploratory visualization". In: *Coordinated and Multiple Views in Exploratory Visualization, 2007. CMV'07. Fifth International Conference on* (2007), pp. 61–71.

[SK95] J. P. Steidlmayer and G. Kummel. *Financial data event flow analysis system with study conductor display*. US Patent 5,454,104. Sept. 1995.

[Spe+16] S. Speakman et al. „Penalized fast subset scanning". In: *Journal of Computational and Graphical Statistics* 25.2 (2016), pp. 382–404.

[TK08] T. Tekušová and J. Kohlhammer. „Visual analysis and exploration of complex corporate shareholder networks". In: *Electronic Imaging 2008*. International Society for Optics and Photonics. 2008, 68090F–68090F.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

[Wan+18] H. Wang et al. „A Visual Analytics Framework for Spatiotemporal Trade Network Analysis“. In: *IEEE transactions on visualization and computer graphics* 25.1 (2018), pp. 331–341.

[Whi11] M. Whitehorn. *The parable of the beer and diapers.* media 2011.

[WWK00] M. Q. Wang Baldonado, A. Woodruff, and A. Kuchinsky. „Guidelines for using multiple views in information visualization“. In: *Proceedings of the working conference on Advanced visual interfaces.* ACM. 2000, pp. 110–119.

[ZC16] B. Zhou and F. Chen. „Graph-structured sparse optimization for connected subgraph detection“. In: *Data Mining (ICDM), 2016 IEEE 16th International Conference on.* IEEE. 2016, pp. 709–718.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

CHAPTER 4

Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration

This paper appeared in: Visual Informatics, 2020. ©Elsevier. Reused with Permission (See [Lei+20])

Full List of Authors Besides Roger A. Leite, the following authors contributed and provided valuable input to this paper: Alessio Arleo, Johannes Sorger, Theresia Gschwandtner, and Silvia Miksch.

4.1 Abstract

The economy of a country can be modeled as a complex system in which several players buy and sell goods from each other. By analyzing the investment flows, it is possible to reconstruct the supply chain for the production of most goods, whose understanding is important to analysts and public officials interested in creating and evaluating strategies for informed and strategic decision making, for instance, adjusting tax policies. Those networks of players and investments, however, tend to be complex and very dense, which leads to over-plotted visualizations that obfuscate precious information such as the dependencies between productive sectors and regions. In this paper, we propose *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*, a guidance-enriched Visual Analytics environment (named after the Greek God of Commerce) for the exploration of complex economic networks, to uncover supply chains, regions' productivity, and sector-to-sector relationships. With practical knowledge regarding guidance, we designed and implemented a visual sub-graph querying approach to extract patterns from

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

such complex investment graphs obtained from real-world data. We present a three-fold evaluation of the system: we perform a qualitative evaluation of our approach with three domain experts, a separate assessment of the proposed guidance features with an expert researcher in this field, and a case study of *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* using a bank account network dataset to demonstrate the generalizability of our approach.

4.2 Introduction

Understanding the behavior of the players in an economic network is a challenging and profitable problem. The obtained knowledge could be used to maintain a sustainable environment to support the country's growth. For example, understanding how in a country productive sectors and regions are related allows the government to plan strategic policies that support positive economic trends. Typical tasks involve looking for weak/strong connections, supply dependencies, or growing and shrinking investment trends. While Artificial Intelligence (AI) is very useful to detect known patterns as well as expected and non-expected outliers, when it comes to reasoning and exploration of a data set, a Visual Analytics (VA) tool works as a brush to an artist. By coupling computational power with the power of the human visual system, we enable domain experts to derive insights which would not be possible with other approaches.

In this work, we focus on the problem of exploring country-wide complex financial networks, aiming at unveiling investment networks, supply chains, and economical links between productive sectors and geographical regions. We present *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*, a VA exploration environment for the exploration of economic networks featuring **time-oriented** and **multivariate** data. Our system is composed of an "Overview" board to provide a high level analysis of the available data and a "Network" board for details on demand. In the latter we propose a visual approach for the definition of patterns to be searched within the network that can be defined by creating and combining "profiles", that are groups of nodes that comply to a set of user-defined constraints. We also include guidance-enriched [Cen+17; CGM19] components to aid the users in the definition of queries and profiles, to enable a fully interactive visual approach to profile creation, pattern drawing, and pattern matching. Therefore, our main contributions are:

- We designed a **visual support for the definition of designated profiles**, which can summarize multiple node types and facilitate a compact construction of query subgraphs.
- We integrated a **guidance-enriched visual subgraph querying approach** that aids the analyst in building subgraph queries from the overabundance of possible node and edge combinations.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- We conducted a **three-fold evaluation**: (1) a task-based qualitative user study with three domain experts collecting insights and guidance update rate measurements, (2) the assessment of our guidance design by an expert in guidance methods for VA, and (3) a use case description demonstrating the generalizability and applicability of *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* to a second domain.

4.3 Related Work

Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration can be framed as a VA approach for the analysis of economic networks using a visual language to perform graph queries with guidance elements. Therefore, we describe the related work by grouping in four categories.

VA for Financial Domain: Visualization proved to be prolific in the financial domain. FinanceVis [DML14] presents a survey involving over 50 research articles, totaling over 70 different techniques, related to the presentation and visualization of financial data. It also features an online collection of the surveyed papers for easy access. The paper concludes that visualization could empower analysts in the financial domain, and encourages further collaboration. Ko et al. [Ko+16] focused on gathering expert feedback and opinions on existing VA approaches in finance. Interviews were also targeted at understanding possible improvements and preferences about visual abstractions, interaction methods, automated techniques, and data sources. When designing *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*, we drew inspiration from the several visualization approaches in the financial domain [DML14], and from the strong motivation and appeal for more real-world based solutions [Ko+16].

Network Analysis in Finance: To enable the VA process on financial networks, graph drawing and visualization played a major role, with node-link representations being often used as an intuitive way to present such networks to the users. Allen and Babus [AB09] argue that node-link graph diagrams are a natural solution for the representation of financial systems and efficiently explain certain economic phenomena such as resources exchange. This directly supports part of our approach design choices. Tekušová et al [TK08] visually supports the users in understanding complex shareholding networks. It incorporates different economic metrics and supports the identification of ultimate shareholders. The identification of network players is a common ground of this work and *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* (see Sec. 4.4.3): R4. VisFAN [Did+11] deals with the observation financial transactions between entities to identify money laundering patterns and potential frauds. The system also allows to cluster and aggregate elements for bottom-up exploration with both automatic and manual clustering techniques. Different from this paper, *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* can group elements by the profile creation feature (see Sec. 4.5.3: A). Huang et al. [HLN09] present an event monitoring approach for stock markets security. It features a linked-views approach

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

with node-link diagrams for network analysis and a 3D treemap for market performance visualization. The main goal of this approach is to identify the attackers (the sources of the fraud), and further attack plans, while reducing the amount of false-positives generated by AI techniques.

Origraph [Big+19] supports network data creation, reshaping, and filtering. The tool allows the user to execute these operations with little to no programming according to the level of details required by the task. In *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*, we were inspired by this work to provide for users with a non-programming approach to generate queries. Differently from Origraph, however, *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* allows to aggregate filters together, to generate profiles that fulfill different constraints. Fujiwara et al. [FCM18] presented a solution to automatically compose a network visual summary based on a rank model (V2D2) and a reduction algorithm. The rank model orders the importance of interactions and provides dependencies among them. Even lacking user studies, the proposal automatic approach proved to be efficient for preparing large networks for analysis. Focused on non-expert users, Elzen et al. [VV14] navigates from network detail to overview by network elements selection and aggregation. In *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*, we follow a similar approach, providing users with different boards that provide different perspectives of the data. As a limitation, they claim to present visual clutter issues with large selections (above 10) due to its clustering or community detection algorithms.

Graphical Queries on Graphs: A typical task in the analysis of financial network is to search for pre-defined patterns and subgraphs, to highlight some transfers and behaviors of interest (e.g., money laundering, tax evasion, supply chain management). To allow users to easily define these patterns, research on VA for finance branched into systems that incorporated visual languages allowing the definition of patterns and profiles of interest, which are used to query the original, complex network. An example of this approach is provided by Didimo et al. [Did+18]. By using an already established visual language for pattern definition [DGM15], they design a system for highlighting malicious transactions between taxpayers (that includes both individuals and legal entities). Differently from our approach, the system is designed to operate in an application domain, while we propose *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* as a more flexible approach more oriented on investments rather than individual transactions. Koenig et al. [KZA10] discuss about the visualization of graph query search results in attributed graphs. This means that each node belongs to a specific category derived by its attributes (e.g., movies, actors, directors). Similarly to our approach, the users define the patterns by combining the different node categories (represented as nodes) together. To display the results, the obtained nodes are mapped to the user-defined patterns, which helps users to keep their mental maps. On the other hand, this approach limits the number of data items that can be shown in the node-link diagram of query results. The problem of designing a suitable visual query language for graph datasets was also discussed by Chau et al. in their GRAPHITE [Cha+08] system and, in more

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

depth, by Pienta et al., with their VISAGE [Pie+16]. Other than providing a platform and visualization for querying graphs, the paper introduces a graph "auto-complete" feature to aid the user to construct and refine queries. This inspired us to include in *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* specific features to assist the users in building the queries. To do so, we focused on the use of "Guidance", a relatively modern field of research that has been recently characterized for VA [Cen+17].

OntoVis [SME06] allows the analysis of social networks by semantic and structural abstraction. It aims at understanding who are the main actors and what are the common behaviors of a social network. Structural abstraction is achieved by the topological information such as node degree and connectivity while semantic abstraction is achieved by filtering nodes that belong to the same type. Different from the "profiles" presented by *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*, that allows for fine-tuned group filters involving different numerical and categorical attributes, the semantic abstraction presented by OntoVis is limited by the existing node types.

Guidance: Guidance has roots in the field of Human Computer Interaction (HCI) [Hor99], but Ceneda et al. [Cen+17] extended van Wijk's model of visualization to integrate guidance into the VA process. They use existing guidance approaches from the literature to illustrate the proposed model. Ceneda et al. [Cen+18a; Cen+18b] designed and implemented a data-driven guidance tool to demonstrate that guidance techniques have the potential to improve data analysis. Most recently, Ceneda et al. [CGM19] conducted an elaborated literature research to provide an overview of how guidance is tackled in different approaches. In this work, the authors highlighted the importance of knowing the actual state of the analysis. We designed *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*' guidance features with the same "context-awareness" principle.

4.4 Design Context

The main goal of this section is to present the design context of *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*, as well as to describe design and implementation choices.

4.4.1 Data, User, and Tasks

We designed and developed our approach based on the design triangle methodology [MA14] and Munzner's nested model [Mun09], by characterizing the problem at hand according to the underlying data, the involved users, and the tasks they have to carry out.

Data. The economic data that our target users analyze can be separated into two categories: the first one describes aggregate information, e.g., on national level, that report the monetary in & out flows between industry sectors within a fiscal year ("IO Tables" [Tim+15]). A transaction in such a table describes, for example, the amount

TU Bibliothek Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar. The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

of investments (in millions of Euros) that went from the "Agriculture" sector to the "Manufacturing" sector. Sectors are encoded according to the ÖNACE specification [ÖNA]. The second data category complements the aggregate information of IO Tables with more granular/regional details, e.g., on the level of individual firms within a region of a country. For this category we used the "Sabina" dataset [Wir]: it contains the balance sheets of all the firms in Austria, including cash flows, locations, branches, and sectors. The combination of both datasets (see Sec. 4.5.1) creates the **Exchange Supply Chain (ESC)** dataset which was used during the design and implementation of *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*. The dataset features 172 nodes and 20,709 edges.

Users. *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* target users are financial and economical experts involved in the development of regional policies and investment plans, for instance in order to economically revitalize a region. Users are expected to be proficient in statistical evaluation of tabular data, but are not expected to have any experience in visual analysis or network analysis.

Tasks. The task that *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* aims at supporting, is to improve the understanding of national economic investment networks based on the integration of aggregated and granular data. After discussing with our domain experts collaborators (not involved on the evaluation), we could identified four sub-tasks: the exploration of the economic flow network existing between sectors (**T1**) and regions (**T2**), the evaluation of how sectors relate and impact each region (**T3**), and supply chain exploration (**T4**) (e.g., predominantly buying/selling regions). The identification of these four sub-tasks support us to better understand the potential of our collaboration with the domain experts.

The combination of the acquired knowledge about data, user, and tasks support us to define the Research Challenges 4.4.2.

4.4.2 Research Challenges

In the context of *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*, we are dealing with several challenges that result from the underlying data as well as the tasks that should be enabled on them. Understanding the research challenges is an important step on the solution design. It is a broader view of the problem that helps us to plan, measure efforts, derive more specific requirements (see Sec. 4.4.3), and derive future work suggestion.

Dense Dynamic Networks: Large economic networks are dense and dynamic, with large and frequent fluctuations. While automatic approaches can still be used to detect outliers and confirm known patterns, VA techniques support economic networks analysis, as they enable users to interactively drill down on the dense data with multiple views, improving the exploration process.

Multi-scale Context: the analysis and exploration of transactions on and across

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

different scales, i.e., regional, national, global, requires visual and analytical scalable solutions.

Geo-political Context: To better understand transactions, we need to consider the motivations that guide them. It is known that geopolitical, geographical, and economical contexts influence this behaviour, e.g., buying from a local or a distant supplier, or the presence of a quarry or oil field. Considering the ever changing local and global scenarios, an analysis tool must be flexible in conveying these different contexts.

Temporal Trends: Economic network analysis involves the observation of stable behaviour over time (e.g., distinguishing growing regional sectors from shrinking and stable ones). However, there are many aspects (scale, scope, arrangement, viewpoint) concerning time-oriented analysis that need to be analyzed efficiently due to their complex data features [Aig+11].

4.4.3 Requirements

By combining the tasks derived by the problem understanding (see Sec. 4.4.1), the broader research challenges context (see Sec. 4.4.2), and with further collaboration of three economists, we derived four main requirements for our system. The economists that collaborated during the requirements and the design development were not participants of the evaluation presented in Section 4.6. By defining a set of requirements, we can also give a better justification for our design choices and provide a robust framework to design our evaluation.

R1: Overview of region and sector distributions. An economic environment is in constant evolution: the transactions between its major players (the companies) can leave a significant and lasting impact on the economy of the region they belong to as well as their foundation and demise. An overview of these aspects is important to understand the main sectors of an economy and supports insights such as the most and least prominent sectors among different regions.

R2: Multi-Scale network exploration. Economists have to understand the national investment and supply chain network to better argue about possible investments, propose tax remodulations, and other types of economic policies. The system must enable network exploration on the different levels of detail on which the information is available. Moreover, the system must implement filter techniques to support the exploration of such a dense network.

R3: Managing queries. During network exploration, analysts regard different aspects of the data by adjusting various filter settings. Often it is necessary to compare and cross-reference the results of such queries. A competent system must thus support the creation, visualization, and comparison of such query profiles.

R4: Identification and tracking of monetary flows. Understanding the monetary flows between sectors and regions could give a significant advantage in many scenarios, including investment planning, supply chain exploration, and fraud detection. The system

TU Bibliothek WIEN Your knowledge hub
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

should support the users in the identification and tracking of the flows in the transactions network, uncovering paths and patterns pertinent to the application domain.

4.5 Design and Implementation

We designed several prototypes that we used to explore the design space in terms of visualization approaches and interactions methods. During different phases of the project development, the prototypes were inspected and evaluated by our three economists collaborators that supported the requirements definition (see Sec. 4.4.3). Our approach was developed using web technology (HTML, CSS, Javascript), as well as frameworks, services, and libraries such as Angular, JQuery, and D3.js. A sketch of the system architecture can be seen in Fig. 4.1.

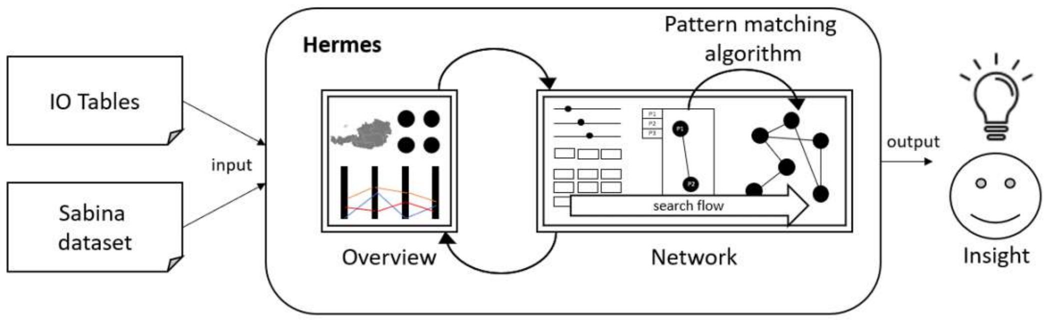

Figure 4.1: *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* workflow. Using the two types of input data (left-hand side) *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* presents two connected visualization boards (Overview and Network), which support interactive exploration and insights generations (right-hand side).

4.5.1 Data Model

Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration data model is represented as a network, whose nodes, called **View Nodes**, represent a productive sector of a specific region (e.g., "Agriculture" in region "Vorarlberg"). Nodes also have attributes, including cash income flow, outcome flow, and growth rate. We obtain the growth rate by calculating the difference of the cashflow (income flow + outcome flow) between the current year (y) and the last year (y-1). Nodes are linked if a financial connection (e.g., investment) exists. We obtain this model from our data sources (see Sec. 4.4) as follows: based on the data combination described by Fritz et al. [Fri+03], we link the details about national companies (i.e., expenditures, amount of employees, sectors, locations) to the monetary income and outcome flow of each productive sector

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

of different regions, enabling us to estimate how much money was exchanged between the different sectors and regions. We can, thus, define different *profiles* to filter these nodes. A *profile* is a set of constraints that identifies a subset of nodes. These constraints might include a group of sectors, a group of regions, in- and outflow ranges, as well as a growing rate range.

4.5.2 Pattern Matching Algorithm

To create graph patterns, we define and combine specific profiles. These profiles are represented as profile nodes and the user can link them using the Network Board (see Sec. 4.5.3). Once a pattern has been constructed, we can search for matching subgraphs in the data model. We define $G_t = (V_t, E_t)$ as our data model graph (i.e., the target graph) described in Sec. 4.5.1, with V_t the vertex set and E_t the set of ordered pairs of nodes representing the graph edges. We also define $|V_t| = n$. $G_p = (V_p, E_p)$ is the pattern the user constructs with profile nodes (i.e., the pattern graph). We use our pattern matching algorithm to look for a subgraph $G_r = (V_r \subseteq V_t, E_r \subseteq E_t)$, i.e. the result graph. This algorithm conducts the following operations: since G_p nodes represent profiles, which identify a set of nodes in V_t, first (**OP1**) for every profile node $v_p \in V_p$ it extracts the nodes $v_t \in V_t$ that match the constraints defined in v_p and assign them to V_r; (**OP2**) for every node now in V_r, it checks whether its outgoing edges belong to E_r. An edge $e_t = (u_t, v_t) \in E_t$ belongs to E_r if $u_t, v_t \in V_r$ and they were extracted from two different profile nodes in V_p connected by an edge $e_p \in E_p$.

It's immediate to show that our algorithm has a quadratic asymptotic time complexity. *OP1* has to be repeated for each node in V_p, therefore has a complexity of $O(|V_p|)$. We can perform *OP2* in $O(d)$ time for each node, with d the out-degree of a node (i.e., the number of outgoing edges). Therefore, considering that from V_p we extract the nodes to insert in V_r, the time complexity of the algorithm is $O(|V_r|d)$: $|V_r|$ is at most n, when the queried subgraph includes all the nodes in V_t, and d is at most $n-1$, when v_r shares an edge with all the other vertices in the graph. Therefore, the asymptotic complexity is $O(n^2)$. In practice, we could find that also with very dense graphs our implementation was able to provide the requested subgraph in very short time, thus enabling an interactive exploration (see Sec. 4.7). We would like to remark that even if the general pattern matching problem is NP-hard, in our case the problem is greatly simplified because the mapping between the profile nodes and the nodes in the target graph is already known, since is defined by the user.

4.5.3 Different Boards for Different Tasks

In this section we describe our design of two combinations of different visualizations (i.e., "boards") serving different tasks. In both boards, we used "Color Brewer" [HB03] to support distinct color pallets for representing different regions. The "multiple boards" approach supports "R2: Multi-Scale network exploration". Both boards are connected by buttons that allow the navigation between each other. It is possible to go from the

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Overview Board to the Network Board by clicking on the right bottom button "Go To Investigation Mode". To go from the Network Board to the Overview Board, the user can click on the "Go To Overview Mode" button in the bottom center. Active filters from one board are carried to the other to keep the same data investigation context.

Overview Board

We designed a fully connected multiple coordinated view board (see Fig. 4.2) to provide a good overview of region and sector distributions (R1) and to support business contextualization. Views are connected with brushing and linking and filter operations in one view are reflected in all the other views accordingly. Besides providing a good overview, this board is designed to support reasoning about how regions are composed of sectors and how sectors are represented in the different regions. Next, we explain our design choices and the possible interactions for each of the views.

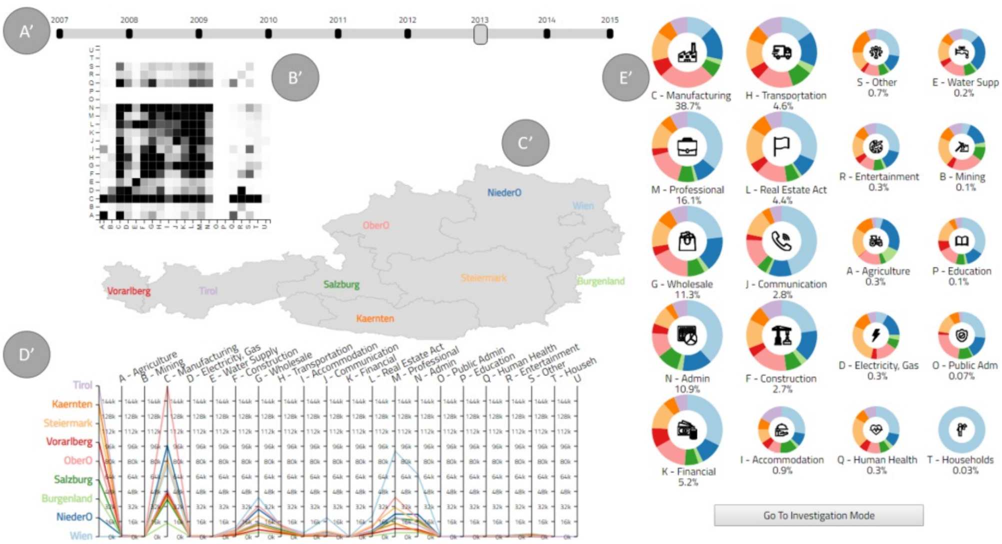

Figure 4.2: Screenshot of *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*: Overview Board. (A') Time Slider: allows the selection of different years. (B') Connection Matrix: gives first impressions about how intense the sectors are connected on a national level. (C') Regions Map: shows the color code and the geographical location of the different national regions. (D') Parallel Coordinates: support users to understand how sectors are distributed among regions. (E') Sectors Donut Charts: shows how regions are distributed among sectors.

A': Time Slider. The main goal of the time slider is to select data from different years for analysis. The user can drag and drop the slider on any year to load the data of this year.

B': Connection Matrix. We show a preview of connections between sectors for the whole country by using an adjacency matrix. We choose adjacency matrices to visualize a

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

dense network [GFC04]. Adjacency matrices also have the advantage that they emphasize an additional variable (i.e., in our case the amount of money transacted between two sectors) instead of connections and paths as node-link diagrams do. Moreover when compared to node-link diagrams, adjacency matrices make better use of the space available and they present less visual clutter. Each square represents a connection and the shade of the square encodes the amount of money transacted. Dark areas in the matrix can be spotted immediately which represent big quantities of money transacted. We also implemented a mouse over feature that displays who is the source and target for each matrix connection square, to make precise information about connections between nodes [KEC06] and amount of money involved available.

C': Regions Map. The Austria national map including different regions not only works interactively by selecting different region's selectors directly in the map, but also supports the contextualization of physically neighboring regions. Therefore, it also supports presentation purposes.

D': Parallel Coordinates. Each colored line represents a different region while the axes represent the amount of companies of each sector. To support different overview tasks, it is also possible to represent other features in the axes such as the total revenue. With this view it is possible to observe how regions are composed by sectors and how sectors are distributed between regions. We opt for using parallel coordinates due to the fact that it supports the representation of multiple dimensions. The axes share a common scale to support comparison.

E': Sectors Donut Charts. A donut chart is essentially a pie chart without the center area. We used the central area for plotting the symbol of each sector. Below each chart, we plot the title of the sector and the percentage that represents the portion of companies within this sector with respect to the all companies in the country (in the selected year). While donut charts should not be used for precise comparisons, we chose them for two reasons (1) to create an area dedicated to easy sector overview and selection (comparable with what view Fig. 4.2: C') does for regions) and (2) to get an immediate impression of the main regions a sector is located in. Also, their aspect ratio allows for the representation of multiple sectors at compact space. In this view, we distinguish between two sizes of donut charts, smaller charts represent sectors that had less than one percent of the national cash flow while bigger charts represents sectors with more than that. This feature facilitates the task of identifying the main sectors of the country. It is possible to select values (sectors and/or regions) from the Overview Board to filter the data that is available in the Network Board, which implements the overview + details on demand mantra [CKB09].

Network Board

During this subsection we refer to view labels from the Fig. 4.3. The Network Board is divided into views A, B, and C to support a three step workflow. In view A, the user creates profiles of sectors and/or regions with specific income and outcome properties as

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

well as growing rates. For example, a profile could be "sectors from Burgenland with a cash income of more than 30 and less than 100 million Euros". In view B, these profiles can then be used as network nodes to construct a network pattern to query the data. The result of a query is displayed in view C. Views D and E present different exploration features that support view C. The workflow presented by the Network Board aims to support "R3: Managing queries" and "R4: Identification and tracking of monetary flows".

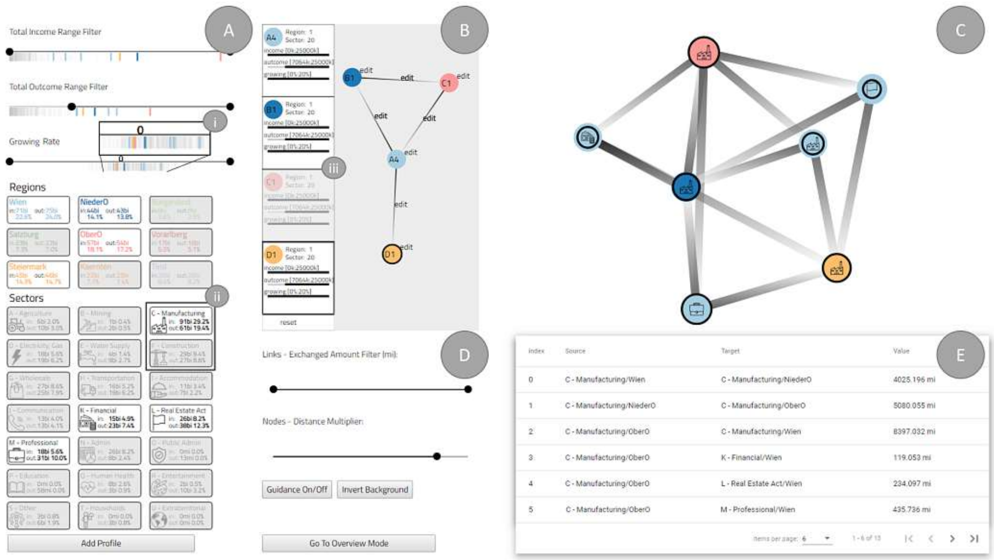

Figure 4.3: Screenshot of *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*: Network Board. (A) Profile Creation View: Profiles created in this view are used to query network patterns in view (B). View (A) is divided into two parts: (i) guidance-enriched sliders and (ii) region/sector filter selection. (B) Guidance-Enriched Pattern Search: here the user can link together profiles to create patterns to be matched. The query result is displayed in (C). (D) Control Panel: this panel allows users to regulate some features from view (C). (E) Dynamic Table: this view shows raw transaction data.

A: Profile Creation View. The profile creation view is divided into two parts: (1) the guidance-enriched sliders to filter income, outcome, and growing rates, and (2) the guidance-enriched tiles to select regions and sectors. Inspired by the Scented Widget [WHA07], the guidance-enriched sliders are designed to increase user familiarity with the data during the profile creation process. Three sliders are presented: income cash flow, outcome cash flow, and growth rate. The income and outcome cash flow dimensions filter how much money a possible sector or region handles while the growing rate filters the growing (or shrinking) rates of sectors and regions. Growing rates are derived by comparing the analyzed year (y) to the previous year (y-1). The guidance aspects to facilitate informed filtering are described in Sec. 4.5.4. Defined profiles are available in view (B) and can be selected as a node to construct a network pattern. Unique profile IDs are assigned which contain a unique letter + the amount of represented regions/sectors.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Profile definitions might be very broad as well as very specific. For instance, a profile which filters only the income dimension to more than 12 billion Euros will match three nodes in the data set described in Sec. 4.4.1. On the other hand, if we filter to less than 12 billion Euros, 160 nodes would match the profile.

B: Pattern Search View. In the Guidance-Enriched Pattern Search View, we have (1) a list of created profiles on the left and (2) a drawing canvas on the right. This view allows users to query for network patterns by selecting profile nodes in (1) and drawing connections between them in (2). A visual summary of the created profiles is shown in (1): a circle filled by the regions' color (or gray if multiple regions are involved) and the profile's ID together with small symbol representing the intervals of income, outcome, and growing rate for this profile. The drawing canvas (2) allows the user visually to recognize the query patterns efficiently [War12]. It supports various profile connection combinations. The canvas also allows for interactively deleting and adding profile nodes and edges as well as changing the direction of edges. Edges might be uni-directional or bi-directional. The generation of profile nodes allows the user to create a compact representation of complex constraints, i.e., multiple sectors and regions with specific characteristics can be combined into one node, which facilitates the construction of network structures to query for and improves clarity of these structures. We present the combination of profiles in (2) as a node-link diagram not only to preserve visual analogy between the constructed pattern and the network data represented in (C), but also to better support path-findings tasks [OJK18]. Moreover, compared to other visualization of network data, such as adjacency matrices, node-link diagrams are better suited to communicate and memorize network topology and connectivity [OJK18]. These tasks are essential for analyzing economic flow.

C: Node-Link View. We opted for a node-link view instead of adjacency matrix due to two main design advantages: (1) it supports visual matching between the pattern drawn by the user in view B and the query results and (2) it supports better reasoning about indirect paths between two nodes (R2 and R4) [KEC06]. The compact network pattern created in view (B) is used to query the data set for sub graphs that match this pattern. The results are then shown in the Node-Link View. However, this view shows a faithful representation of the data set without collapsing multiple sectors into one node, as it is done in view (B) to facilitate pattern construction. Thus, each node in this view represents a specific sector of a specific region and edges represent the yearly money flow between them. The size of nodes is proportional to the amount of money they handle (income & outcome) while the ring close to the outer border of the node, indicates the grow rate by representing the size of the node in the previous year (see Fig. 4.4). In other words, if the ring of a node is bigger than the node itself, the amount of money handled by this sector was decreasing since last year. On the other hand, if the ring is smaller than the node, this sector has a growing trend (see Fig. 4.4). The amount of money exchanged between sectors is encoded by edge width. To define this width, we again use a logarithmic scale to avoid extreme visual outliers. Following Holen and van Wijk's [HV09] recommendations, we avoided standard arrow representation and, instead,

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

chose a dark-to-light encoding. The money flows from the node at the dark side of the edge to the node at the light side. For bi-direction edges both sides are dark (see leftmost node in Figure 4.3). All nodes and edges can be selected for viewing detail information in the Dynamic Table View (see view (E)).

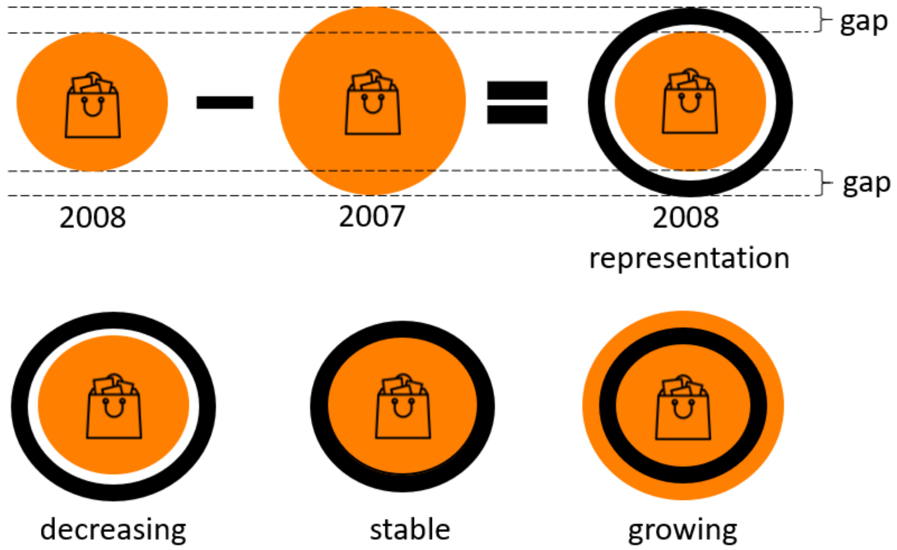

Figure 4.4: Visual encoding of the growing rate of a sector. The first row shows the comparison operation and how it is reflected in the visual encoding of a sector node. Using this metaphor, we can visually represent shrinking, stable, and growing sectors (second row).

D: Control Panel. The control panel was implemented to facilitate the inspection and refinement of the network presented in view (C). The first slider allows filtering edges by amount of money and the second slider allows re-scaling the network by increasing or decreasing the forces between the nodes. Since edges directions are represented by a dark-to-light gradient, we provide the option to invert the the background of the node-link diagram to black, which facilitates the identification of receiver patterns.

E: Dynamic Table. During our problem analysis together with experts, we learned that they prefer to rely on detailed information presented in a spreadsheet-like representation. While tables are not suited to provide a good overview of the network structure, they are good in presenting all the detail information of specific sectors and regions. Thus, we implemented a dynamic table to provide experts with these details on demand. The dynamic table is populated with information about all edges visible in view (C). Selecting edges in this table highlights their representations in view (C).

4.5.4 Guidance Aspects

In this subsection we describe the different degrees and features of guidance that are provided by *Hermes: Guidance-enriched Visual Analytics for Economic Network Explo-*

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

ration. Guidance by itself is a computer-assisted process that continuously reduces the knowledge gap that may block the data exploration and analysis. It supports the users to make sense of the data on their own [Cen+17; Cen+18a; CGM19].

A huge number of profile combinations and constructed network patterns would be feasible, but only a small set of them make sense for the data at hand. To avoid frustrating trial and error interaction until a meaningful pattern for querying the data set could be found, we needed to design *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* with features that guide the user in this pattern construction. *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* includes four different features aimed at providing guidance to the user, and thus, facilitating a meaningful and informed usage of the tool. The following indexes "i,ii,iii" refer to Fig. 4.3 and "iv" refers to Fig. 4.5.

i: Slider Ticks. Sliders in the Profile Creation View are augmented with ticks colored with respect to the different regions (see Fig. 4.3: i). One tick represents one sector in one region. Hovering these ticks makes a tool tip window appear which shows full textual information about the sector, region, income, outcome, and growing rate. Ticks are grayed out when they do not match all defined filters. Thus, this visualization of ticks at the point of interaction (i.e., the slider) guides the user in making a meaningful selection of sectors for profile construction, by a priory avoiding constructions that would not lead to any results as these constructs cannot be found in the particular data set. This kind of guidance (i.e., adding information that facilitates the interaction and selection of the data) meets the guidance degree of orienting [Cen+17] which is aimed at building and preserving a user's mental map [AP12].

ii: Selection Tiles. The tiles for selecting regions and sectors work in accordance with the slider filters, for including or excluding regions and/or sectors from the filtered selection. Options that, in combination with the other filter settings, do not exist in the data are grayed out (see Fig. 4.3: ii). This, again, helps to avoid the construction of non-existing profiles. Since this feature limits the users interaction possibilities (i.e., grayed-out tiles cannot be selected), it corresponds to the guidance degree of directing, which usually offer the user a pre-selected pool of alternative next steps for a given task [Cen+17].

iii: Guided Network Pattern Construction. While the two previous guidance features help the user to define meaningful profiles, the combination of these profile nodes into a network pattern that can be found in the data, is also a laborious task if performed without guidance. Thus, the Pattern Search View (see Sec. 4.5.3 (B)) does not allow for the construction of non-existing patterns. This is achieved in a two-step process: First, every time the user selects a profile node that is part of the constructed pattern in the drawing canvas (in Fig. 4.3: B) in order to add a connection to another profile node, the system runs a query that looks for possible connections from this node in the data set, and disables the selection of all profile nodes that would lead to empty query results. For example, a given "profile C" that would lead to empty query results when connected with "profile A". However, it might be possible to connect it with "profile B". In Fig. 4.3: iii,

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

the node D1 is selected in the pattern drawing canvas and thus, profile C1 is disabled in the list of available profiles by the system. Again, this feature limits users in their possibilities to construct arbitrary patterns, which makes it directing guidance [Cen+17].

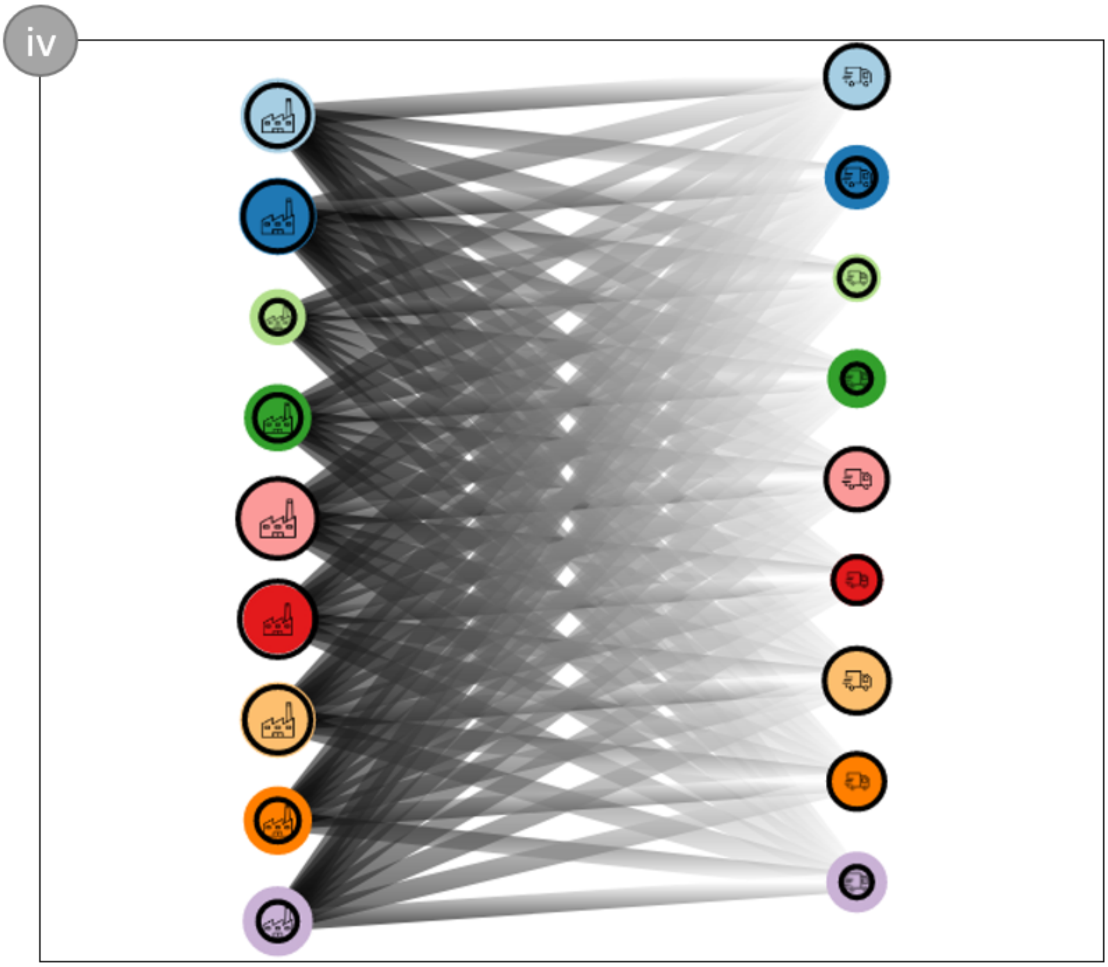

Figure 4.5: In this picture, we see a bipartite layout result of two profiles. On the left-hand side, the first profile is representing the manufacturing of each region and, on the right-hand side, the second profile represents the transportation sector of each region. Based on this result, we can investigate how much each manufacturing region spends with each transportation region.

iv: Switching to Bipartite Layout. Another design choice that helps orienting the user [Cen+17], is that the network presentation in the Node-Link View switches to a bipartite layout (see Fig. 4.5) every time the user constructs a profile combination in the Pattern Search View that connects only two profiles. The automatic adaption of the layout creates an additional awareness in the user that he/she is now investigating a bipartite graph. While this is obvious from the query pattern constructed in the pattern

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

canvas, it can be hard to spot in the Node-Link View which nodes from the query result belong to which profile. Moreover, the layout facilitates the flow analysis between the two groups.

4.6 Evaluation

To assess and evaluate the effectiveness and limitations of *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*, we conducted a qualitative user study, in the form of an interview with three domain experts (P1, P2, P3) [Ise+13; KP15]. We first describe the user study design (see Sec. 4.6.1) followed by the participant background (see Sec. 4.6.2); afterwards we discuss the tasks and the immediate findings of our study. To complement our study, we interviewed a research expert in guidance (P4), who discussed his/her thoughts about the theoretical and practical implications of the guidance features of *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* (see Sec. 4.6.5). Furthermore, we also did examination of the transferability of *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* to Financial Fraud Analysis (FFA) with a domain expert (P5).

4.6.1 Study Design

We conducted our user study as a task-based evaluation: the objective was to evaluate and observe how users operate *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* when asked to solve various tasks. Our study design consists of individual interviews followed by a hands-on session to be carried out by up to two participants at a time. Each interview is structured as follows: (1) participants background interview, (2) introduction of overview and Network Board, (3) hands-on testing, (4) task-based evaluation, and (5) final general feedback discussion. Each session lasted about 90 minutes. We aimed at making the users familiar with *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* before asking them to perform the task-based evaluation individually and (mostly) unguided. For the whole duration of the session, participants' interaction with *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* was recorded. They were also asked to think-aloud in order to follow their reasoning during the evaluation. The tasks chosen are different for the overview and Network Boards and were designed to evaluate whether *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* could fulfill its design requirements (see Sec. 4.4.3). We identified four tasks per board (see Sec.s 4.6.3 and 4.6.4), structured according to the analytical task taxonomy by Andrienko and Andrienko [AA06], which highlights differences between elementary and synoptic tasks.

Furthermore, to support insight analysis, we categorize the obtained user insights by following Klein's model of intelligence analysis [Kle13]. Our taxonomy includes: **Connection** insights, obtained by combining data from two or more views; **Coincidence** insights, which include relationships between elements or events that are obscured at first, but that the users can perceive as connected; **Curiosity** insights happen when

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

users do not find information when they expect to, and this pushes them to find a reason why. Finally, in order to fully comprehend the impact of our guidance techniques in our study, we implemented in our prototype a counter that would keep track of how many times the system updated the interface to guide the users (see Sec. 4.5.4). We called this measurement **Guidance Updates**. The number of insights and guidance updates are displayed in Fig. 4.6 for each task. During the evaluation session with P1, P2, P3, and P4, we used *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* with the *ESC* dataset (see Sec. 4.5.1). The FFA dataset was exceptionally presented to P5, who is domain expert in the field (see Sec. 4.6.6).

4.6.2 Participants' Background

P1 and P2: The participating experts (not involved in the design) can all be considered potential *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* target users (see Sec. 4.4.1). The two participants are employees from the economic and trade policy department of the "Austrian Federal Economic Chamber" (*Wirtschaftskammer Österreich* or WKO). They analyze the trends of business interests in Austria as well as evaluating and predicting the effects economic policies and regulations throughout the different regions of the country. Both participants use Microsoft Excel [Mic] as well as "Python" and "R" to explore and present their data. However, they report a lack of visual tools that support their work. Therefore, it is expected that they also lack visual guidance. One participant was familiar with "Tableau", however does not apply it in his work context.

P3: Our third participant is an economist, who works as the global treasury controller within an international company. In contrast to the other two participants, who looked after the productivity of entire regions and multiple companies, P3 is interested in the company's business only. Therefore, P3 is looking for a system that could support investment decisions with information about product supply chain networks. The most used tool by P3 is Microsoft Excel Sheets [Mic] and P3 claims to be familiar with simple visual data representations existing o it.

P4: Aiming for a focused qualitative evaluation of the guidance aspects in *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*, we invited a VA expert researcher with a research focus on guidance. The participant was used to develop VA techniques.

P5: Our fifth participant is a financial fraud expert that works for a national bank. P5 claim being familiar with a good range of visualization techniques and, on her/his work, s/he uses visualizations for presentation and exploration proposes.

4.6.3 Evaluation of the Overview Board (P1, P2, P3)

We designed the Overview Tasks (**OT**) to evaluate the first of the two *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*'s boards, which was designed to fulfill requirement R1 (see Sec. 4.4.3).

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

The tasks are the following:

OT1: Which are the biggest sectors in Austria?

OT2: Which regions present more manufacturing companies?

OT3: What is the main region? Why?

OT4: By looking and interacting to the Matrix View (Fig. 4.2: B'), what type of conclusions could you take?

All the participants concluded the tasks before the specified time, and reported correct answers. Connection insights were the most common insight category during the evaluation of the Overview Board. However, OT4 did not present any connection insight because it is an one view task. Overall, we interpret this result as a positive consequence of presenting a fully interactive multiple coordinate view approach that shows different dimensions of the same data elements. The combination of different selection and filtering features was the main aspect explored during the performance of OT1, OT2, and OT3.

OT3 was especially designed to motivate value thinking by presenting the words "main region" on its specification without describing what was the comparison context encapsulated by the word "main". The participants could define different meanings to the word "main". Still the three participants could explore their hypothesis and reasoning about the different possible outcomes of this interpretation, thus, revealing the flexibility of the tool.

4.6.4 Evaluation of the Network Board (P1, P2, P3)

For this board we also conceived four tasks. To evaluate R3 (see Sec. 4.4.3) we designed two similar network tasks (NT1 and NT2). The goal is understanding the impact of guidance during the profile construction phase. To do so, we deactivated the guidance features in *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* when evaluating NT1 and reactivated them on NT2. In addition, we asked our participants to tackle two open tasks (NT3 and NT4), which allowed for multiple correct asnwers. The goal is to observe how many insights and conclusions the participants could draw within the given 5 minute time frame. NT3 and NT4 aim at supporting R2 and R4 analysis (see Sec. 4.4.3).

NT1: Excluding the region "Oberösterreich", which View Nodes are the sectors with the four highest *income*? How are they connected? The task has to be completed without guidance support.

NT2: Excluding the region "Oberösterreich", which View Nodes are the sectors with the four highest *outcome*? How are they connected? In this task the guidance features are enabled.

NT3: Choose two regions and explore their strongest connections. Discuss your findings.

NT4: Choose a region and search for growing sectors and shrinking sectors. Discuss your findings.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Without guidance support, two out of the three experts were not able to finish NT1. P2 achieved it but past the allotted 5 minutes mark. Without visual guidance during profile creation, participants were not able to precisely set the income slider to a convenient income range, that led to the inclusion of a large amount of nodes (noise) in each of the created profile nodes. As a consequence, the resulting matched subgraphs did not help comprehending the data and solving the task. On the other hand, the attempts at narrowing the income range failed at including the minimum amount of searched nodes specified by NT1.

With guidance support (NT2), all participants were able to finish the similar task within less than two minutes. Guidance thereby facilitates the determination of a threshold that incorporates data points relevant to the task. By observing the colored ticks under the outcome slider, users could understand how many nodes they were selecting, thus setting a "convenient" slider range. Moreover, since the excluded region is color coded on the Slider Ticks, the participants could exclude samples from the selection without necessarily interacting with selection region filters (see Sec. 4.5.3), thus speeding up the profile creation process. Tasks NT3 and NT4 are open exploration tasks that were designed to motivate open insights. Both tasks resulted in the most significant amount of insights across all tasks (see Fig. 4.6).

Overall, the participants provided positive feedback about *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* and complemented different aspects: *"The guidance features are really useful"* (P1) ,*"... compared to the technology I use, to perform network exploration tasks, this approach is much faster and appealing"* (P3), or *"...to perform the same tasks with our current approaches is almost impossible"* (P2). P3 also commented positively about the possibility to switch to a bipartite layout: *"...it is good when I search for less groups because the layout gets more organized"*. All participants gave positive feedback on the temporal aspect derived by the growing rate slider. It presented a main role during many of the participants' insights elaboration.

Considering the feedback and the results of the task evaluation, the data appears to support our claims, confirming that *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* fulfills the design requirements. NT1/2 show the capabilities of *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* when handling queries, with the guidance support (R3). The amount of findings in NT3/4 and the feedback gathered suggest that *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* succeeds in enabling multi-scale network exploration (R2) and the tracking of monetary flows and region/sector connection (R4).

4.6.5 Guidance Expert Feedback (P4)

The focus of this evaluation was to have P4 to interact with the tool to collect expert feedback about the guidance aspects and not measure insights or answers precision. After performing the complete evaluation, during the post tasks interview, the participant mentioned several positive points. S/He complemented different aspects such as the

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

short interaction response time of the Overview Board and the tasks enabled by the Network Board. S/He highlighted the importance of the guidance features during the profile creation (see Sec. 4.5.3 view A). Starting with the Slider Ticks (see Fig. 4.3: i) s/he commented: "*...with the guidance it works perfectly to select the profile that you want to see, otherwise it is complicated to find what you want. It gives you hints about the data that you are selecting*". Concerning "Selection Tiles" (see Fig. 4.3: ii) and "Guided Network Pattern Construction" (see Fig. 4.3: iii), P4 considered this guidance aspect as something in between the *orienting degree* and the *directing degree*, more inclined to directing degree (see Sec. 4.5.4). Concerning the "Switching to Bipartite Layout" (see Fig. 4.5: iv), s/he commented: "*it is a good choice that the system supports this automatic decision. It is visually clear for me that this is a better layout for comparison tasks and group understanding*".

When asked about future recommendation for guidance features, a point raised by our participant was the limits of guidance (see Sec. 4.7). How much guidance is enough? S/He commented that guidance should always serve a propose, a task. However, when trying to add guidance to many tasks in the same tool, guidance features trend to overcome each other, leading users to confusion. Based on that comment, s/he mentioned that all

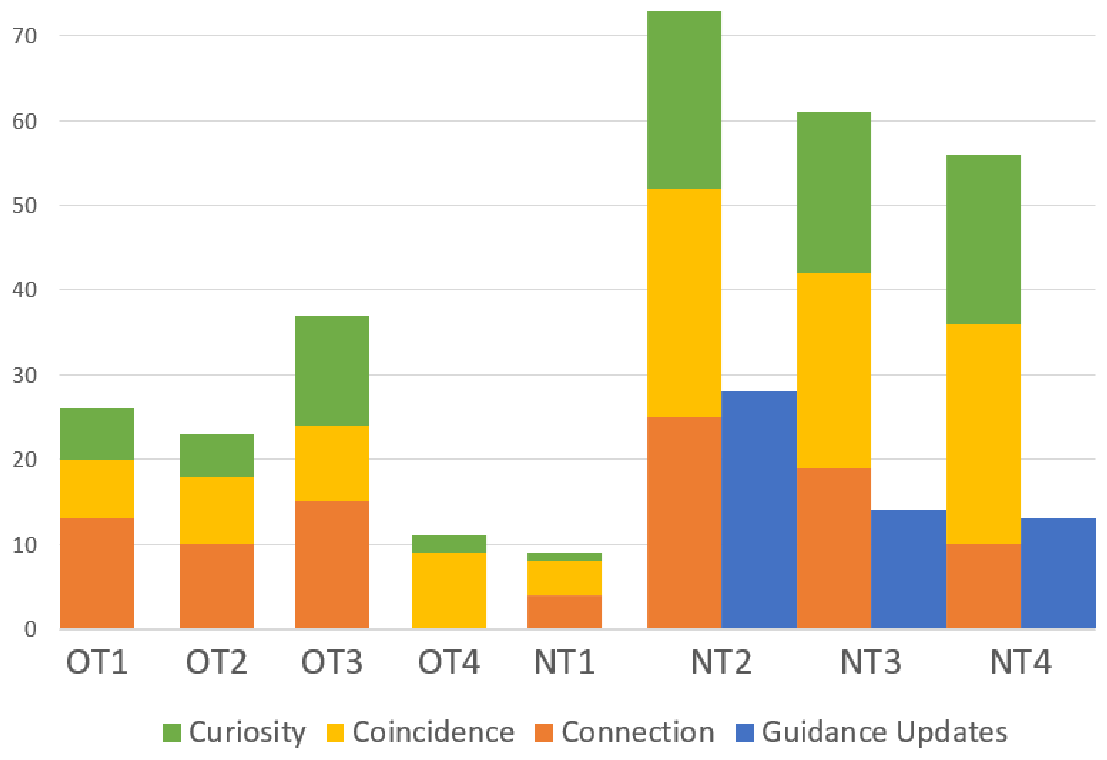

Figure 4.6: The overall number of and guidance updates we identified over the different board tasks. We report both the exact numbers in a table and an area chart to easily extract trends from it.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

proposed tasks that were covered by guidance aspects, were not disturbing each other and, s/he would not suggest any further guidance implementation, unless the system is extended to support new tasks. In the same sense, s/he also complemented that the profile creation view does not enforce a specific workflow, avoiding influencing users to limit their profile creation: *"...it is good to avoid steps repetition during exploration tasks"*.

Overall, s/he was surprised to see different degrees of guidance in *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*. Her/His favorite features were the guidance implemented in the profile creation and the pattern search view. S/He highlighted several times how much easier and better it was to have guidance features supporting her/him during these tasks.

4.6.6 Applying *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* to Bank Account Networks (P5)

We demonstrate the *generalizability* of our approach by applying it to a different domain, i.e., the analysis of financial fraud. The FFA data is a real-world dataset for which we render bank accounts as nodes and money transaction between accounts as edges between them. Transaction are ranked according to a metric that measures its how likely it is to be part of a fraudulent behaviour calculated by a bank internal automatic algorithm. *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* could be used by fraud investigators for tasks such as identifying terrorists, bank credit exploration, or fraud analysis. For these cases, profiles of bank accounts could be composed of features like age, balance, transaction frequency, account growing rate. With the combination of profile definition and profile-based network pattern queries the investigators would gain a supportive and flexible ally.

We also had the opportunity to present *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* applied to FFA data to a FFA expert (P5), and to collect positive feedback. P5 was able to interact and understand all *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*' features. In particular, the expert appreciated: (1) the potential of flexible profile construction coupled with the network pattern search for detecting networks of fraudsters, (2) the quick pattern search response of *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*, and (3) the guidance aspects during profile creation - which s/he assessed essential for understanding what s/he was selecting and allowing for an informed profile creation.

4.7 Discussion and Future Work

In this section we list the limitations of *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* and we discuss some general aspects about guidance that

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

came up during the discussions with the guidance expert. Moreover, we derive directions for future work.

Limited Support of Profile Membership Identification. We provide a feature that changes the layout of the node-link diagram to help the user understand which nodes belong to which profiles (see Sec. 4.5.4: iv). By now this feature is limited to bipartite graphs (i.e., including only two profiles). Because this feature was very well received in the evaluation, we plan to incorporate a similar functionality also for queries including more than two profiles. This comes with additional challenges of how to arrange the graph layout when the number of profiles exceeds a certain limit. Radial layouts, for instance, might be implemented to address this problem.

Connections Between Profiles are AND Operations From our requirement analysis we derived the need of analysts to search for patterns where all node types of one profile are connected to all node types of another profile. Thus, connecting two profiles in *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* queries the data for network structures in which nodes of different profiles with an AND (i.e., all nodes of one profile are connected to all nodes from the other profile). However, new ideas came up in the evaluation and thus, we want to further investigate if connecting two profiles with an OR operation would provide additional benefits to analysts.

Pre-defined Profiles & Network Structures. Another idea stemming from our evaluation with domain experts is to provide pre-defined profiles as well as pre-defined network structures of common queries. For example, experts suggested a profile that involves only elements with a high grow rate or a profile of elements with a high outcome. They also suggested that pre-defined network structures would support an even faster analysis. However, to provide pre-defined profiles and network structures, domain knowledge is required. Moreover, it is not clear how such "ready to use" building blocks influence users in their analysis and if they would limit exploration breadths [Wal+17].

Performance Considerations. As we described in Sec. 4.4.1, we experimented with the ESC dataset, a fairly dense graph. Despite this, thanks to the simplicity of our approach (see Sec. 4.5.2), we observed that *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* featured low response times also with more complex queries. During our expert evaluation and internal testing, we tested queries that included stars (a central node connected to some neighbors), cycles, disconnected components, and combinations thereof up to 21 profile nodes (whose constraints identified several nodes in the model graph). We measured the response times, from the pattern extraction to the display of the matched sub-graph, which peaked at 67 milliseconds.

Classification of Guidance. It is not always unambiguous how to classify the provided guidance and even what should be considered as guidance. However, the expert agreed that quite often features might present a mixture of different degrees. For instance, all Selection Tiles that do not exist in the data in combination with the given filter settings are grayed out. This might be considered just a hint for orientation (i.e., orienting guidance), but it also presents users with a list of meaningful alternatives to proceed

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

with, which is considered directing guidance. Sometimes basic visualization features comply with the definition of guidance. For example, in a node-link diagram, encoding the connectivity degree of a node in the node's size to emphasize how well a node is connected might also be considered some sort of orienting guidance. However, these basic features are common practice in the visualization community and, therefore, are not considered to be guidance. Further research is needed to better distinguish guidance.

4.8 Conclusion

We developed *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration*, a guidance-enriched VA approach to support national economic supply chain analysis on an overview level as well as on a detailed network analysis level. The design and implementation of *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* was inspired by our own research experience and by tight collaboration with economy domain experts, with whom we elaborated a number of design requirements. Our approach consists of two connected boards which provide overview and network views augmented with various guidance aspects to support economy experts with effectively exploring and analyzing complex data.

We designed and evaluated *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* with two real-world data sets and five domain experts (three economists, one FFA expert, and one guidance researcher expert), from whom we obtained positive feedback and further ideas on various aspects of our design. The added value of our approach is: (1) the analysis of insights gained by our evaluators, (2) a positive evaluation of the guidance aspects by a guidance researcher expert, and (3) the demonstration of the transferability of *Hermes: Guidance-enriched Visual Analytics for Economic Network Exploration* to FFA.

4.9 Acknowledgments

This work was partially supported by the Research Cluster "Smart Communities and Technologies (SmartCT)" at TU Wien and the Austrian Science Fund (FWF), grant P31419-N31 (Knowledge–Assisted Visual Analytics (KnoVA).

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Bibliography

[AA06] N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. ISBN: 978-3-540-31190-4. DOI: 10.1007/3-540-31190-4_3. URL: http://dx.doi.org/10.1007/3-540-31190-4_3.

[AB09] F. Allen and A. Babus. „Networks in finance“. In: *The network challenge: strategy, profit, and risk in an interlinked world* 367 (2009).

[Aig+11] W. Aigner et al. *Visualization of time-oriented data*. Springer Science & Business Media, 2011.

[AP12] D. Archambault and H. C. Purchase. „Mental map preservation helps user orientation in dynamic graphs“. In: *International Symposium on Graph Drawing*. Springer. 2012, pp. 475–486.

[Big+19] A. Bigelow et al. „Origraph: interactive network wrangling“. In: *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE. 2019, pp. 81–92.

[Cen+17] D. Ceneda et al. „Characterizing guidance in visual analytics“. In: *IEEE Transactions on Visualization & Computer Graphics* 1 (2017), pp. 111–120.

[Cen+18a] D. Ceneda et al. „Guidance or No Guidance? A Decision Tree Can Help“. In: *EuroVis Workshop on Visual Analytics (EuroVA). The Eurographics Association*. 2018.

[Cen+18b] D. Ceneda et al. „Guided Visual Exploration of Cyclical Patterns in Time-series“. In: *Visualization in Data Science*. Posterpräsentation: Visualization in Data Science (VDS at IEEE VIS 2018), Berlin; 2018-10-22. IEEE Digital Library, 2018.

[CGM19] D. Ceneda, T. Gschwandtner, and S. Miksch. „A review of guidance approaches in visual data analysis: A multifocal perspective“. In: *Computer Graphics Forum* 38.3 (2019), pp. 861–879.

[Cha+08] D. H. Chau et al. „Graphite: A visual query system for large graphs“. In: *Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on*. IEEE. 2008, pp. 963–966.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

[CKB09] A. Cockburn, A. Karlson, and B. B. Bederson. „A review of overview+ detail, zooming, and focus+ context interfaces". In: *ACM Computing Surveys (CSUR)* 41.1 (2009), p. 2.

[DGM15] W. Didimo, F. Giacchè, and F. Montecchiani. „Kojaph: visual definition and exploration of patterns in graph databases". In: *International Symposium on Graph Drawing and Network Visualization*. Springer. 2015, pp. 272–278.

[Did+11] W. Didimo et al. „An advanced network visualization system for financial crime detection". In: *Pacific Visualization Symposium (PacificVis), 2011 IEEE*. IEEE. 2011, pp. 203–210.

[Did+18] W. Didimo et al. „A visual analytics system to support tax evasion discovery". In: *Decision Support Systems* 110 (2018), pp. 71–83.

[DML14] M. Dumas, M. J. McGuffin, and V. L. Lemieux. „Financevis.net - A Visual Survey of Financial Data Visualizations". In: *Poster Abstracts of IEEE VIS 2014*. Poster and Extended Abstract. Nov. 2014.

[FCM18] T. Fujiwara, T. Crnovrsanin, and K.-L. Ma. „Concise provenance of interactive network analysis". In: *Visual Informatics* 2.4 (2018), pp. 213–224.

[Fri+03] O. Fritz et al. „Constructing Regional Input-Output Tables for Austria". In: *Austrian Economic Quarterly* 1 (2003), pp. 23–39.

[GFC04] M. Ghoniem, J.-D. Fekete, and P. Castagliola. „A comparison of the readability of graphs using node-link and matrix-based representations". In: *IEEE Symposium on Information Visualization*. IEEE. 2004, pp. 17–24.

[HB03] M. Harrower and C. A. Brewer. „ColorBrewer. org: an online tool for selecting colour schemes for maps". In: *The Cartographic Journal* 40.1 (2003), pp. 27–37.

[HLN09] M. L. Huang, J. Liang, and Q. V. Nguyen. „A visualization approach for frauds detection in financial market". In: *Information Visualisation, 13th International Conference*. IEEE. 2009, pp. 197–202.

[Hor99] E. Horvitz. „Principles of mixed-initiative user interfaces". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM. 1999, pp. 159–166.

[HV09] D. Holten and J. J. Van Wijk. „A user study on visualizing directed edges in graphs". In: *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM. 2009, pp. 2299–2308.

[Ise+13] T. Isenberg et al. „A systematic review on the practice of evaluating visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2818–2827.

[KEC06] R. Keller, C. M. Eckert, and P. J. Clarkson. „Matrices or node-link diagrams: which visual representation is better for visualising connectivity models?" In: *Information Visualization* 5.1 (2006), pp. 62–76.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

[Kle13] G. Klein. *Seeing what others don't: The remarkable ways we gain insights.* PublicAffairs, 2013.

[Ko+16] S. Ko et al. „A Survey on Visual Analysis Approaches for Financial Data“. In: *Computer Graphics Forum*. Vol. 35. 3. Wiley Online Library. 2016, pp. 599–617.

[KP15] S. Kriglstein and M. Pohl. „Choosing the Right Sample? Experiences of Selecting Participants for Visualization Evaluation“. In: *EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (EuroRV3)*. Ed. by W. Aigner, P. Rosenthal, and C. Scheidegger. The Eurographics Association, 2015. DOI: 10.2312/eurorv3.20151146.

[KZA10] P.-Y. Koenig, F. Zaidi, and D. Archambault. „Interactive searching and visualization of patterns in attributed graphs“. In: *Proceedings of Graphics Interface 2010*. Canadian Information Processing Society. 2010, pp. 113–120.

[Lei+20] R. A. Leite et al. „Hermes: Guidance-enriched Visual Analytics for economic network exploration“. In: *Visual Informatics* 4.4 (2020), pp. 11–22.

[MA14] S. Miksch and W. Aigner. „A matter of time: Applying a data–users–tasks design triangle to visual analytics of time-oriented data“. In: *Computers & Graphics* 38 (2014), pp. 286–290.

[Mic] Microsoft. *Excel*. office.microsoft.com/en-us/excel/ (accessed: 2016-12-09).

[Mun09] T. Munzner. „A nested model for visualization design and validation“. In: *IEEE transactions on visualization and computer graphics* 15.6 (2009), pp. 921–928.

[OJK18] M. Okoe, R. Jianu, and S. G. Kobourov. „Node-link or Adjacency Matrices: Old Question, New Insights“. In: *IEEE transactions on visualization and computer graphics* (2018).

[ÖNA] ÖNACE. *ÖNACE 2008 - Struktur*. Accessed: 2019-07-20. http://www.statistik.at/KDBWeb/kdb_VersionAuswahl.do.

[Pie+16] R. Pienta et al. „Visage: Interactive visual graph querying“. In: *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM. 2016, pp. 272–279.

[SME06] Z. Shen, K.-L. Ma, and T. Eliassi-Rad. „Visual analysis of large heterogeneous social networks by semantic and structural abstraction“. In: *IEEE Transactions on Visualization and Computer Graphics* 12.6 (2006), pp. 1427–1439.

[Tim+15] M. P. Timmer et al. „An Illustrated User Guide to the World Input–Output Database: the Case of Global Automotive Production“. In: vol. 23. 3. 2015, pp. 575–605.

[TK08] T. Tekušová and J. Kohlhammer. „Visual analysis and exploration of complex corporate shareholder networks“. In: *Electronic Imaging*. International Society for Optics and Photonics. 2008.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

[VV14] S. Van den Elzen and J. J. Van Wijk. „Multivariate network exploration and presentation: From detail to overview via selections and aggregations“. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 2310–2319.

[Wal+17] E. Wall et al. „Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics“. In: *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE. 2017, pp. 104–115.

[War12] C. Ware. *Information visualization: perception for design.* Elsevier, 2012.

[WHA07] W. Willett, J. Heer, and M. Agrawala. „Scented widgets: Improving navigation cues with embedded visualizations“. In: *IEEE Transactions on Visualization and Computer Graphics* 13.6 (2007), pp. 1129–1136.

[Wir] Wirtschaftsuniversität Wien. *Sabina - Info - Datenbanken.* Accessed: 2019-06-10. https://www.wu.ac.at/bibliothek/recherche/datenbanken/info/sabina/.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Part III

Conclusion

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

CHAPTER 5

Conclusion

5.1 Research Questions Revisited

In this section, we revisit the research question as described in section 1.4 and answer them according to our publications. We first respond to the sub-questions to, later, support a final answer for the main question. In this section we refer to Chapter 2 as "EVA", Chapter 3 as "NEVA", and Chapter 4 as "Hermes".

Sub-Question 1: How can we address the particular characteristics of users, time-oriented and multivariate data, and tasks?

Time-oriented and multivariate data present different challenges [Aig+11]. We had both data features always present in our approaches designed to tackle problems in different domains. EVA and NEVA share a focus on the Financial Fraud Detection (FFD) domain. EVA is focused on the fine-tuning of the detection algorithm, while NEVA focuses on the relationships within the bank's users network. In both works, time was a central aspect of the problem and was supported by a multiple coordinated views approach. Graphs designed to represent time were connected to different data dimensions' representation to support the proposed tasks. Moreover, all charts were always connected by interactions. For example, brushing a period in the time graph would make elements from the other charts appear or disappear according to the new filtered context. In Hermes, we worked within the economic supply exchange domain. The tasks given by the experts avoided fine-grain analysis of time. For example, to support decision-making at a governmental level, it is good to prevent the misleading potential of high volatility presented by a fine-time grain of the market. Instead, they limited the time analysis granularity to focus on macro changes (yearly). Moreover, the exchanges between different sectors (IO Tables) are estimated officially once a year in the country [Wir], because it is attached to the tax payment information collection.

TU WIEN Bibliothek Your knowledge hub

Every problem resulted in a slightly different VA approach, as expected. Each domain requires different tasks that can be optimized by different data analysis, visualization, and interaction techniques. However, in our works, all solutions shared a "multiple coordinated views" feature responsible for connecting "time" to our "various dimensions". This connection was directly done by different dimensions in the same graph graphs, or indirectly, by interactions that supported responsible views connectivity.

Sub-Question 2: Can we derive an appropriate workflow for supporting different user groups and their tasks?

We showed through our works that an appropriate workflow is derivable to support different user groups and their tasks. EVA, NEVA, and Hermes share similar human-centered workflow. During the design and development of these approaches, a human-centered workflow featured a loop between experts and design decisions. The main idea of these loops is to bring real-world user experience and solution acceptance to the development. Even when the relation between designers and the target audience appeals like a "must have relation", many designers tend to assume task ranks and values without asking the future users about it [Mun09b]. In other words, the loop between designers and real users avoids unnecessary features and fine-tune necessary ones. Having a loop into the workflow improves not only (1) the performance of needed tasks but also supports (2) the discovering of new tasks. Both are essential features of a given solution. Next, we elaborate on both, taking into consideration a human-centered workflow.

The tasks' performance (1) is improved by avoiding the usage of interactions, visualizations, and automatic techniques that are not known by the domain public. It is true that not only with well-known techniques a solution should be made for every problem. When the domain can profit from new techniques, the loop methodology can monitor the potential public's learning curve. In other words, we can derive common mistakes and difficulties from the users. Thus, the design can avoid such situations by using guidance-enriched components or changing it completely to other approaches that can, again, be tested in the loop (up to the point when the tasks and the users are suitable to each other). During the human-centered design loop, (2) discovering new tasks is also a potential exciting outcome. The suggestion of new functional tasks was a widespread event through our works. By getting in touch with prototypes, different users often give hints about enhancing existing task implementation and new tasks that could be useful to the overall solution.

The knowledge and time invested of a domain expert are unreachable by VA designers during the given time period of solution development. Therefore, having a human-centered approach, bringing experts to the development loop, is a must-have feature of a VA design workflow that supports: (1) fine-tune of existing tasks and (2) discovery of new tasks.

Sub-Question 3: How can we generalize our contribution to other domains?

In this cumulative thesis, we presented three works from two different domains. Even EVA and NEVA being from the same FFD domain, it is true that they focus on different

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

sub-domains: Automatic algorithm fine-tuning and customers‘ network analysis. On the other hand, Hermes focuses on the economic supply chain domain. Every new challenge will always be unique in a way. However, there are contributions that we can take from our works to improve the success rate of a new approach. Each distinct VA solution presented in this thesis shared similar design methodology that could be replicated to other solutions from different domains. The distinct VA approaches presented by us are adaptable to other domains by (i) identifying similar problems throughout the domains and (ii) coupling variables with similar features (quantitative, categorical, etc.). For example, EVA presented a scoring system approach that is adaptable to support the fine-tuning of any other automatic algorithm that bases its decision on a final score based on various sub-scores calculations. Another example is the guidance-enriched components for network pattern generation and detection from NEVA and Hermes. In this example, the guidance-enriched components are explored in two different domains (FFD and Economic Supply Exchange). Moreover, they support different analysis complexity levels through these different domains. Therefore, any domain containing network features could profit from such a technique.

VA approaches‘ features are shaped by the combination of automatic algorithms, visualization, and interaction techniques. Within each solution, these three elements are usually fine-tuned to Data-User-Tasks specifications. However, different domains share similar characteristics by their data types and this was the main point of exploration during generalizing features from one application domain (EVA) to the other (Hermes). Moreover, the scientific reference base of design, implementation, and evaluation methodologies achieved in our approaches is replicable to different domains. Furthermore, researches frontiers suggested by our results could also inspire new frontiers in other disciplines.

Main Research Question:

How can VA techniques support event analysis by investigating time-oriented and multivariate data in different domains by accomplishing various tasks and users?

Based on our works, we conclude that (1) VA approaches that tackle temporal and multivariate data by connecting different points of view are an efficient way to improve the speed and quality of data exploration and performance of decision-making tasks. Showing different “angles” of the same dataset is effective when (2) connecting different views with responsive interactions. (3) Guidance components aid users to develop more efficient queries results and data limitation understanding. Therefore, complex data analysis empowered by VA approaches can support tasks regardless the domain.

Another insight was that the knowledge and time invested of a domain expert are unreachable by any VA designers during the given time period of project development. A human-centered approach that brings experts to the development loop, is a must-have feature of a VA project workflow. The loop supports not only (1) fine-tune of existing tasks but also (2) discovery of new tasks.

Every problem solution is unique by its Data-User-Tasks specifications. However, different domains share similar characteristics by their data types and this is the main exploration point to generalize features from one application domain to the other (e.g.,NEVA and Hermes). Moreover, the combination of scientific methodologies adopted to design, implement, and evaluate in our approaches is replicable to different domains. Furthermore, researches frontiers suggested by our results could also inspire new frontiers in various disciplines.

5.2 Discussion and Future Works

This section presents a discussion about the open challenges and opportunities grouped by data and task complexity, visual scalability, multi-coordinated views for interactive exploration, VA approach, and evaluation according to particular domains and tasks. Moreover, we derive open research challenges and directions to inspire future works.

5.2.1 Data and Task Complexity

Event detection and exploration techniques may vary concerning different scientific research domains. One of the reasons for that is the type of data involved. It is not easy to get real-world data that usually belong to companies and contain private interest information. This difficulty is mainly due to privacy and security reasons. Moreover, real-world datasets often are not complete having some features hidden or changed to preserve costumers' privacy [Lei+15a].

When it comes to different application domains, we need to identify and specify different types of events. For example, bank frauds are usually more network-related and include the identification of "payment fraud", "money laundry", and "straw persons". Insurance frauds, on the other hand, often are analyzed by checking if a sequence of events is plausible and if fraudulent patterns exist. Internal in companies fraud detection often uses process mining related solutions. Telecommunication frauds are tackled with rule-based visual systems. Stock market fraud detection uses a broader range of techniques due to the various types of frauds that can be found in this domain.

One aspect that adds up to the complexity of event detection is that finding suited solutions for detecting suspicious cases is not enough [DR15]. In the case of frauds, for instance, new fraud techniques are always upcoming or being re-adapted. Fraudsters can be very creative when it comes to hiding their attempts. One example is to hide attacks in known and non-suspicious patterns of events. This strategy may cause simple rule-based approaches to fail. Besides, another challenge in the field is to find a monitoring solution [HLN09]. The detection of already happened events and the prevention of future similar circumstances is a critical task, and so is the prediction of possible critical events such as frauds. This organic behavior existing in real-world data makes the task of event detection complex and challenging.

Suited solutions need to avoid false-positive identification as much as possible, which would burden the investigators and waste their time of analysis. False-negatives, which

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

miss actual recurrent events and, by consequence, may result in fraudulent harm [Lue10]. In other words, to be more helpful than harmful, the solutions need to be precise in estimating possible threats fine-tuned to each application domain.

To analyze and estimate event threats, interpreting single events does not usually lead investigators to conclusions. However, a sequence of events, or a network of events, allows the investigator to reason about suspicious behavior by comparing historical events within their contexts [Cha+07].

5.2.2 Visual Scalability

In event detection, independent of the application domain, the data is often multivariate, always temporal, and comprises enormous amounts of data items. For example, (1) daily bank transactions of a vast amount of costumers [Lei+15a], (2) a bid and offer a variation of the NASDAQ [HLN09], or (3) the internal operations made by all employees of a company in different systems during a specific period [ASS13] are challenging to be visually represented. This difficulty is partly due to the fact that the data is not only multivariate but usually also covers long periods of time. To this end, visual aggregation techniques are often needed in order to display such rich datasets. However, during analysis, the exploration of individual cases or short period analysis might still be an exciting task. Thus, interaction techniques such as elaboration, exploration, and filtering are usually applied to support these tasks.

5.2.3 Multi-Coordinated Views for Interactive Exploration

Detecting events according to particular criteria within time-oriented and multivariate datasets is a challenging task. This process asks for an intertwined visual and analytical approach in an interactive multiple-coordinated exploration environment. However, many of the states of the art approaches made limited use of interaction to support the analysis task. Some approaches used a loose coupling of views (compare, for example, [Sch+11] and [ASS13]), others a more closed coupling (compare, for instance, WireVis [Cha+07]). Moreover, the most popular visualization methods are line plots, node-link diagrams, and bar charts [Lei+18b]. It is also known that 3D approaches are getting less popular recently [Lei+18b]. Investigating in systematical applicability of various interaction and visualization techniques according to particular tasks would open new possibilities to explore and analyze fraudulent behavior. This challenge is closely related to the next one.

5.2.4 VA Approach

Solely automated methods often fail to detect fraudulent behavior because actors are strategically changing their behavior to mislead monitoring and catching them. This solution asks for a VA approach. To our knowledge, only WireVis [Cha+07] pursues a VA approach. The user (in our case, the investigator) should take an active role in selecting automatic or analytical approaches, fine-tuning the parameter settings,

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek, WIEN Your knowledge hub

interactive exploration of the dataset, etc. and a seamless integration thereof. In other words, there is a lot of open space to support the various steps in the knowledge generation process [Sac+14].

5.2.5 Evaluation According to Particular Domains and Tasks

Different domains and tasks in event detection demand a similar evaluation. Evaluating VA based event detection solutions is difficult due to the lack of experts in each area. Domain knowledge is crucial to perform an event analysis. A typical practice is to select a group of investigators to analyze a new tool/solution and further ask for empirical feedback. At the same time that the evaluation participants must have the background to perform a fraud analysis and he or she should not have previous knowledge concerning the analyzed dataset.

In most articles, the evaluators came from the same companies, banks, or insinuations that provided the dataset. However, to perform a fair evaluation and avoid previous knowledge to influence the results, the investigator should not be too familiar with the datasets. Otherwise, the evaluation participants could be subconsciously influenced to find a determine outlier or pattern that he or she already knew. On the other hand, it is hard to find investigators who have a suitable background to analyze such datasets.

Approaches that address a domain with a high degree of social and financial impact, such as fraud detection, should be carefully evaluated to guarantee the worth of substitution of the already existing approaches and the investment of implementing these new solutions into a real-world system. However, in the surveyed papers, it is common to use a small number of evaluators (between 2 and 5), usually with previous knowledge about the dataset. To measure if an evaluation has enough secure results to justify the solution's further investment or not, is still an open problem. On the other hand, the urgent demands to detect, analyze, and monitor events are continually increasing. Therefore, we are optimistic that more event analysts will formulate their demands and needs. There is still enough research space to conduct qualitative and quantitative evaluations to access the usability and usefulness of the event detection and exploration proposed solutions.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

5.3 Conclusion

This thesis described three different visual analytics techniques for event analysis: EVA, NEVA, and Hermes. Deep analytical and domain knowledge was required to develop this thesis, which considers the financial fraud detection domain and the economic supply exchange domain.

Within our findings, we concluded that three features are core to VA solutions involving time-oriented and multivariate data: (1) multiple coordinates view, (2) responsive interactions, and (3) guidance components. The three systems benefit from similar scientific background for design, implementation, and evaluation. A Data-User-Tasks approach supports the design and implementation of our solutions [MA14b; Mun09a]. We received positive feedback from respective domain experts [Ise+13; KP15] in our works. During the assessments, we measured the insights gained by the systems [Kle13] and confirmed the concrete benefits from our approaches. With EVA, NEVA, and Hermes, we demonstrate that VA techniques improve domain experts' learning curves and insights into data exploration, comprehension, and manipulation. Overall, we experienced that complex data analysis empowered by VA approaches can support different exploration and decision-making tasks regardless of the domain.

Based on our findings, we also propose future research challenges and directions. These challenges cover beyond the domains presented in this thesis (financial fraud detection and economical supply exchange). We showed that VA solutions can be generalized and greatly support different fields with similar problems and data characteristics such as biology, environment, medicine, social sector, insurance, public sector, and many others.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Bibliography

[A L+20] R. A. Leite et al. „NEVA: Visual analytics to identify fraudulent networks“. In: *Computer Graphics Forum*. Vol. 39. 6. Wiley Online Library. 2020, pp. 344–359.

[Aig+11] W. Aigner et al. *Visualization of time-oriented data*. Springer Science & Business Media, 2011.

[AK13] F. Atefeh and W. Khreich. „A survey of techniques for event detection in twitter“. In: *Computational Intelligence* (2013).

[Arl+19] A. Arleo et al. „Sabrina: Modeling and Visualization of Financial Data over Time with Incremental Domain Knowledge“. In: *2019 IEEE Visualization Conference (VIS)*. IEEE. 2019, pp. 51–55.

[ASS13] E. N. Argyriou, A. A. Sotiraki, and A. Symvonis. „Occupational fraud detection through visualization“. In: *2013 IEEE, Intelligence and Security Informatics (ISI)*. IEEE. 2013, pp. 4–6.

[Bög+17] M. Bögl et al. „Cycle Plot Revisited: Multivariate Outlier Detection Using a Distance-Based Abstraction“. In: *Computer Graphics Forum*. Vol. 36. 3. Wiley Online Library. 2017, pp. 227–238.

[CBK09] V. Chandola, A. Banerjee, and V. Kumar. „Anomaly detection: A survey“. In: *ACM computing surveys (CSUR)* 41.3 (2009), pp. 1–58.

[CC08] J. D. Cryer and K.-S. Chan. *Time series analysis: with applications in R*. Springer Science & Business Media, 2008.

[Cha+07] R. Chang et al. „Wirevis: Visualization of categorical, time-varying data from financial transactions“. In: *Visual Analytics Science and Technology. VAST. IEEE Symposium on*. IEEE. 2007, pp. 155–162.

[Cle93] W. S. Cleveland. *Visualizing data*. Hobart Press, 1993.

[DR15] W. N. Dilla and R. L. Raschke. „Data visualization for fraud detection: Practice implications and a call for future research“. In: *International Journal of Accounting Information Systems* 16 (2015), pp. 1–22.

[Edw83] T. Edward R. „The Visual Display of Quantitative Information“. In: (1983).

[Faw+98] T. Fawcett et al. „AI approaches to fraud detection and risk management.“ In: *AI Magazine* 19.2 (1998), p. 107.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

[Fun36] H. G. Funkhouser. „A note on a tenth century graph". In: *Osiris* 1 (1936), pp. 260–262.

[Fun37] H. G. Funkhouser. „Historical development of the graphical representation of statistical data". In: *Osiris* 3 (1937), pp. 269–404.

[Gau+13] D. Gauldie et al. „Louvain Clustering for Big Data Graph Visual Analytics". In: (2013).

[HLN09] M. L. Huang, J. Liang, and Q. V. Nguyen. „A visualization approach for frauds detection in financial market". In: *Information Visualisation, 13th International Conference*. IEEE. 2009, pp. 197–202.

[Ise+13] T. Isenberg et al. „A systematic review on the practice of evaluating visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2818–2827.

[Kei+08] D. Keim et al. „Visual Analytics: Scope and Challenges". In: *Visual Data Mining*. Ed. by S. J. Simoff, M. H. Böhlen, and A. Mazeika. Springer, 2008, pp. 76–90. DOI: 10.1007/978-3-540-71080-6_6.

[Kle13] G. Klein. *Seeing what others don't: The remarkable ways we gain insights*. PublicAffairs, 2013.

[KP15] S. Kriglstein and M. Pohl. „Choosing the Right Sample? Experiences of Selecting Participants for Visualization Evaluation". In: *EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (EuroRV3)*. Ed. by W. Aigner, P. Rosenthal, and C. Scheidegger. The Eurographics Association, 2015. DOI: 10.2312/eurorv3.20151146.

[Lei+15a] R. A. Leite et al. „Visual Analytics for Fraud Detection and Monitoring". In: *Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on*. IEEE. 2015, pp. 201–202.

[Lei+15b] R. A. Leite et al. „Visual analytics for fraud detection and monitoring". In: *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE. 2015, pp. 201–202.

[Lei+16] R. A. Leite et al. „Visual Analytics for Fraud Detection: Focusing on Profile Analysis." In: *EuroVis (Posters)*. 2016, pp. 45–47.

[Lei+17] R. A. Leite et al. „EVA: Visual analytics to identify fraudulent events". In: *IEEE transactions on visualization and computer graphics* 24.1 (2017), pp. 330–339.

[Lei+18a] R. A. Leite et al. „Network Analysis for Financial Fraud Detection." In: *EuroVis (Posters)*. 2018, pp. 21–23.

[Lei+18b] R. A. Leite et al. „Visual analytics for event detection: Focusing on fraud". In: *Visual Informatics (STAR)* 2.4 (2018), pp. 198–212.

[Lei+20a] R. A. Leite et al. „COVIs: Supporting Temporal Visual Analysis of Covid-19 Events Usable in Data-Driven Journalism". In: *2020 IEEE Visualization Conference (VIS)*. IEEE. 2020, pp. 56–60.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[Lei+20b] R. A. Leite et al. „Hermes: Guidance-enriched Visual Analytics for economic network exploration". In: *Visual Informatics* 4.4 (2020), pp. 11–22.

[LF10] S. Lai and H.-C. Fu. „A polygon description based similarity measurement of stock market behavior". In: *Expert Systems with Applications* 37.2 (2010), pp. 1113–1123.

[Lue10] J. Luell. „Employee fraud detection under real world conditions". PhD thesis. University of Zurich, 2010.

[MA14a] S. Miksch and W. Aigner. „A matter of time: Applying a data–users–tasks design triangle to visual analytics of time-oriented data". In: *Computers & Graphics* 38 (2014), pp. 286–290. DOI: `10.1016/j.cag.2013.11.002`.

[MA14b] S. Miksch and W. Aigner. „A matter of time: Applying a data–users–tasks design triangle to visual analytics of time-oriented data". In: *Computers & Graphics* 38 (2014), pp. 286–290.

[May+19] E. Mayr et al. „Visualizing biographical trajectories by historical artifacts: A case study based on the photography collection of Charles W. Cushman". In: *Paper accepted for publication at the Biographical Data in a Digital World Conference*. Vol. 2019. 2019.

[Mun09a] T. Munzner. „A nested model for visualization design and validation". In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009), pp. 921–928.

[Mun09b] T. Munzner. „A nested model for visualization design and validation". In: *IEEE transactions on visualization and computer graphics* 15.6 (2009), pp. 921–928.

[ÖNA] ÖNACE. *ÖNACE 2008 - Struktur*. Accessed: 2019-07-20. `http://www.statistik.at/KDBWeb/kdb_VersionAuswahl.do`.

[RG13] D. Rosenberg and A. Grafton. *Cartographies of time: A history of the timeline*. Princeton Architectural Press, 2013.

[Sac+14] D. Sacha et al. „Knowledge generation model for visual analytics". In: *Visualization and Computer Graphics, IEEE Transactions on* 20.12 (2014), pp. 1604–1613.

[Sak+10] C. Sakoda et al. „Visualization for Assisting Rule Definition Tasks of Credit Card Fraud Detection Systems". In: *IIEEJ Image Electronics and Visual Computing Workshop*. 2010.

[Sal+19] S. M. Salisu et al. „Shapes of Time: Visualizing Set Changes Over Time in Cultural Heritage Collections." In: *EuroVis (Posters)*. 2019, pp. 45–47.

[ŠB10] A. Šilić and B. D. Bašić. „Visualization of text streams: A survey". In: *Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2010, pp. 31–43.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

[Sch+11] M. Schaefer et al. *A Novel Explorative Visualization Tool for Financial Time Series Data Analysis*. Bibliothek der Universität Konstanz, 2011.

[Shn96] B. Shneiderman. „The eyes have it: A task by data type taxonomy for information visualizations“. In: *Proc. of IEEE Symposium on Visual Languages*. IEEE, 1996, pp. 336–343. DOI: 10.1109/VL.1996.545307.

[SLM21] A. Scheidl, R. Leite, and S. Miksch. „VisMiFlow: Visual Analytics to Support Citizen Migration Understanding Over Time and Space“. In: IEEE, 2021.

[TC05] J. J. Thomas and K. A. Cook. *Illuminating the path: The research and development agenda for visual analytics*. Tech. rep. Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2005.

[TG83] E. R. Tufte and P. R. Graves-Morris. *The visual display of quantitative information*. Vol. 2. 9. Graphics press Cheshire, CT, 1983.

[Til75] L. Tilling. „Early experimental graphs“. In: *The British Journal for the History of Science* 8.3 (1975), pp. 193–213.

[Tim+15] M. P. Timmer et al. „An Illustrated User Guide to the World Input–Output Database: the Case of Global Automotive Production“. In: vol. 23. 3. 2015, pp. 575–605.

[TK08] T. Tekušová and J. Kohlhammer. „Visual analysis and exploration of complex corporate shareholder networks“. In: *Electronic Imaging*. International Society for Optics and Photonics. 2008.

[Tuk77] J. W. Tukey. *Exploratory data analysis*. Vol. 2. Reading, Mass., 1977.

[Win+20] F. Windhager et al. „Many Views Are Not Enough: Designing for Synoptic Insights in Cultural Collections“. In: *IEEE Computer Graphics and Applications* 40.3 (2020), pp. 58–71.

[Wir] Wirtschaftsuniversität Wien. *Sabina - Info - Datenbanken*. Accessed: 2019-06-10. https://www.wu.ac.at/bibliothek/recherche/datenbanken/info/sabina/.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

List of Figures

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

List of Tables

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.