Flesch, Benjamin Johannes

Doctoral Thesis

Social Set Visualizer (SoSeVi): Design, development and evaluation of a visual analytics tool for computational set analysis of big social data

PhD Series, No. 22.2019

Provided in Cooperation with:
Copenhagen Business School (CBS)

Suggested Citation: Flesch, Benjamin Johannes (2019) : Social Set Visualizer (SoSeVi): Design, development and evaluation of a visual analytics tool for computational set analysis of big social data, PhD Series, No. 22.2019, ISBN 978-87-93744-87-5, Copenhagen Business School (CBS), Frederiksberg,
https://hdl.handle.net/10398/9740

This Version is available at:
https://hdl.handle.net/10419/222896

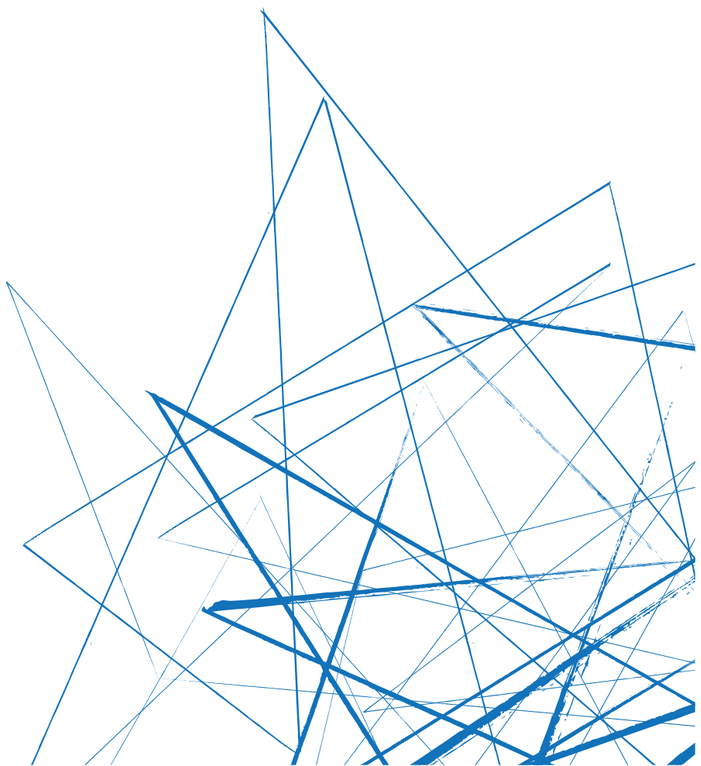
COPENHAGEN BUSINESS SCHOOL
SOLBJERG PLADS 3
DK-2000 FREDERIKSBERG
DANMARK

WWW.CBS.DK

ISSN 0906-6934

Print ISBN: **978-87-93744-86-8**
Online ISBN: **978-87-93744-87-5**

SOCIAL SET VISUALIZER (SOSEVI): DESIGN, DEVELOPMENT AND EVALUATION OF A VISUAL ANALYTICS TOOL FOR COMPUTATIONAL SET ANALYSIS OF BIG SOCIAL DATA



PhD Series 22-2019

Benjamin Johannes Flesch

SOCIAL SET VISUALIZER (SOSEVI)

DESIGN, DEVELOPMENT AND EVALUATION OF A VISUAL ANALYTICS TOOL FOR COMPUTATIONAL SET ANALYSIS OF BIG SOCIAL DATA

Doctoral School of Business and Management | **PhD Series 22.2019**

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

Copenhagen
Business School
HANDELSHØJSKOLEN

Social Set Visualizer (SoSeVi):

Design, Development and Evaluation of a Visual Analytics Tool for Computational Set Analysis of Big Social Data

Benjamin Johannes Flesch

A PhD dissertation presented to the faculty of the
Doctoral School of Business and Management at
Copenhagen Business School

Supervisors

Prof. Ravi Vatrapu	Copenhagen Business School
Prof. Raghava Rao Mukkamala	Copenhagen Business School

Copenhagen Business School June 2019

Benjamin Johannes Flesch
Social Set Visualizer (SoSeVi): Design, Development and Evaluation of a Visual Analytics Tool for Computational Set Analysis of Big Social Data

1st edition 2019
PhD Series 22.2019

ISSN 0906-6934
Print ISBN: 978-87-93744-86-8
Online ISBN: 978-87-93744-87-5

The Doctoral School of Business and Management is an active national and international research environment at CBS for research degree students who deal with economics and management at business, industry and country level in a theoretical and empirical manner.

Acknowledgements

This PhD project would not have been possible without the help of numerous kind people who have helped me along the way and carried me with their good spirits.

First of all, I want to thank **Ravi Vatrapu**, my primary supervisor. Through his high energy and positive aura he always finds solutions where others see problems. Furthermore, he catalyzes an outstandingly warm and productive working atmosphere at the Centre for Business Data Analytics. He gave me the opportunity to become part of his team and collaborate on a large variety of interesting research projects over the past few years. Thank you very much.

I want to thank **Raghava Rao Mukkamala**, my secondary supervisor, with whom I have spent numerous hours tinkering about code, conceptual models, and teaching data science, and who has always given me insightful comments and knowledgeable advice. Thanks for your kindness and resourcefulness.

My PhD project was financially supported by **Industriens Fond**. Therefore, I want to thank all stakeholders at the Danish Industry Foundation for supporting this journey.

Many deep-felt thanks to **Niels Buus Lassen**, my great friend and fellow PhD student. We had many good times in Copenhagen and beyond, yet you always let me crash on your couch if times were tough. Thank you so much.

Thanks also to the many new friends I was lucky to make during this journey. Many thanks to **Abid Hussain**, to **Lester Allan Lasrado**, to **René Madsen**, to **Helgi Waag**, to **Kjeld Hansen**, to **Nadia Straton**, to **Cancan Wang**, to **Jakob Berg Jespersen**, to **Povl Gad**, to **Kalina Staykova**, to **Tor-Morten Grønli**, to **Chris Zimmerman**, to **Christian Casper Hofma**, to **Juan Giraldo**, to **Mikkel Harlev**, to **Nikolina Vukelic**, to **Nikola Ens**, to **Elham Shafiei**, to **Peter Wynne**, to **Kim Normann Andersen**, to **Katrine Kunst**, to **Henning Langberg**, and to all dear colleagues at the department that I have forgotten. I am very thankful for your positive influence during my time as a PhD student.

Thank you to **Kiran Kocherla** and **Dharanidaran Aladiyur Paramisivan** for your very pleasant companionship as researchers in our lab. Thank you to **Bodil Sponholtz**, **Jeanette Hansen** and **Cecilie Ostenfeld** for your great support in any and all administrative matters.

Lastly, my deepest thanks to my family and my girlfriend, without whose support this would not have been possible.

Abstract

This dissertation presents the design, development and evaluation of the Social Set Visualizer, an innovative Visual Analytics software tool, that expands upon a novel set-based approach to Big Social Data Analytics for large-scale datasets from social media platforms such as Facebook. Over the course of five peer-reviewed publications, three different versions of the Visual Analytics software tool are iteratively designed and developed, and several contributions to the visualization of sets and set intersections are highlighted.

In seven case studies with the Social Set Visualizer software tool the generation of meaningful facts and actionable insights from Big Social Data are empirically demonstrated, and a pre-existing research gap with regard to the Visual Analytics of large-scale Facebook datasets vs. other social media platforms is closed. Based on these studies, the dissertation puts forward a generalized conceptual model for interactions within Big Social Data termed the Social Interaction Model, which provides a simplification and extension of previous theoretical and formal models.

. . . . . .

Denne afhandling præsenterer design, udvikling og evaluering af Social Set Visualizer, et innovativt Visual Analytics softwareværktøj, der udvider sig på en ny setbaseret tilgang til Big Social Data Analytics til store datasæt fra sociale medieplatforme som Facebook. I løbet af fem peer-reviewed publikationer er tre forskellige versioner af Visual Analytics software værktøjet iterativt designet og udviklet, og flere bidrag til visualisering af sæt og sæt kryds er fremhævet. I syv cases-

tudier med softwareværktøjet Social Set Visualizer er genereringen af meningsfuldt fakta og brugbare indsigter fra Big Social Data demonstreret empirisk og et eksisterende forskningsgab med hensyn til Visual Analytics af store Facebook-datasæt versus andre sociale Medieplatforme er lukket. På baggrund af disse studier fremlægges afhandlingen en generel konceptuel model for interaktioner inden for store sociale data, der betegnes social interaktionsmodellen, som giver en forenkling og forlængelse af tidligere teoretiske og formelle modeller.

Table of Contents

List of Acronyms

ADR	Action Design Research
API	Application Programming Interface
AR	Augmented Reality
BDVA	Big Data Visual Analytics
BSD	Big Social Data
BSDA	Big Social Data Analytics
CSR	Corporate Social Responsibility
CSS	Computational Social Science
CSV	Comma-Separated Values
DSR	Design Science Research
GIS	Geospatial Information System
GUI	Graphical User Interface
HCI	Human-Computer Interaction
MR	Mixed Reality
RDBMS	Relational Database Management System
SDM	Social Data Model
SIM	Social Interaction Model
SNA	Social Network Analysis
SoDaTo	Social Data Tool
SoSeVi	Social Set Visualizer
SQL	Structured Query Language
SSA	Social Set Analysis
SSQL	Social Set Query Language
SVG	Scalable Vector Graphics
UCD	User-Centric Design
UI	User Interface
UTF	Unicode Transformation Format
UX	User Experience
VA	Visual Analytics
VR	Virtual Reality

List of Tables

List of Listings

List of Figures

Chapter 1

Introduction

This PhD project contributes to the advancement of the state of the art in the domain of Computational Social Sciences by **providing two novel solutions to the key challenges** of *"working with different data formats and structures"* and *"developing methods for visualizing massive data"* identified in the National Academy of Sciences' report on massive data analysis [National Research Council *et al.* 2013].

First, this PhD project addresses the challenge of ***"working with different data formats and structures"*** in the domain of Computational Social Science by proposing the Social Interaction Model, a formal model of social interactions that is agnostic to the technical aspects of social media data from sources such as Facebook, Twitter, Instagram, WeChat, or Sina Weibo.

Second, this PhD project addresses the challenge of ***"developing methods for visualizing massive data"*** from the domain of Visual Analytics by interactively visualizing large-scale sets and set intersections in the Social Set Visualizer (SoSeVi). The Social Set Visualizer provides a novel way of insight generation, using Social Set Analysis, the set-theoretical approach to Big Social Data Analytics that was pioneered by our research group at the Centre for Business Data Analytics. Specifically, this PhD project tackles the challenge that arises when large-scale sets calculated from Big Social Data need to be accurately visualized and analyzed. It is resolved through application of innovative set visualization techniques to Big Social Data Analytics, which in turn enable users of the Social Set Visualizer software tool to utilize the Social Set Analysis approach to its full potential.

In order to facilitate the understanding of the key challenges identified above and the PhD project's contributions, this introductory chapter will first establish the foundations of Big Social Data Analytics and then outline the research problem.

1.1 Background

In recent years, **Big Data** ermerged as a term describing the increasing volumes of data which are difficult to store, process, and analyze through traditional database technologies and analytical means [Hashem *et al.* 2015]. For Big Data, a variety of definitions exist. Most prominently, the 3V definition of Big Data, based on volume, velocity, and variety, originally devised by Gartner, and the 4V definition, based on volume, velocity, variety, and veracity [Gantz & Reinsel 2011], are used.

Due to the increasing availability of Big Data, the challenge of performing **Big Data Analytics** with the ambition to discover meaningful facts and actionable insights has risen to utmost importance both in industry and academia [Wamba *et al.* 2017].

For the purposes of this thesis, Big Data Analytics is defined as *"a set of techniques and technologies that require new forms of integration to uncover hidden values from large datasets that are diverse, complex, and of a massive scale"* [Hashem *et al.* 2015].

Big Data created through the widespread use of social media has been termed **Big Social Data** [Bello-Orgaz *et al.* 2016,Vatrapu *et al.* 2016] and is defined as *"high-volume, high-velocity, high-variety and highly semantic data that is generated from technology-mediated social interactions and actions in the digital realm; which can be collected and analyzed to model social interactions and behavior"* [Olshannikova *et al.* 2017].

For the research presented in this PhD project, Facebook represents the most important source of Big Social Data with Twitter, Instagram, YouTube, and Reddit as additional data sources of relevance. Based on various research projects utilizing both Big Data and Big Social Data, considerable differences between Big Data Analytics and **Big Social Data Analytics** have been identified. This distinction is based on significant differences in sources and structure of data, as well as in social diversity and cultural relativity, the analytical focus on symbolic and textual components of social interactions, and the strong emphasis on security, privacy, and ethics [Vatrapu *et al.*]. Therefore, we argue that Big Social Data Analytics should be seen as a distinct subfield of its own within Big Data Analytics.

Having outlined the distinction betwen Big Data Analytics and Big Social Data Analytics, two important concepts are now introduced which build the basis for the overarching Big Social Data Analytics research framework employed in this PhD project.

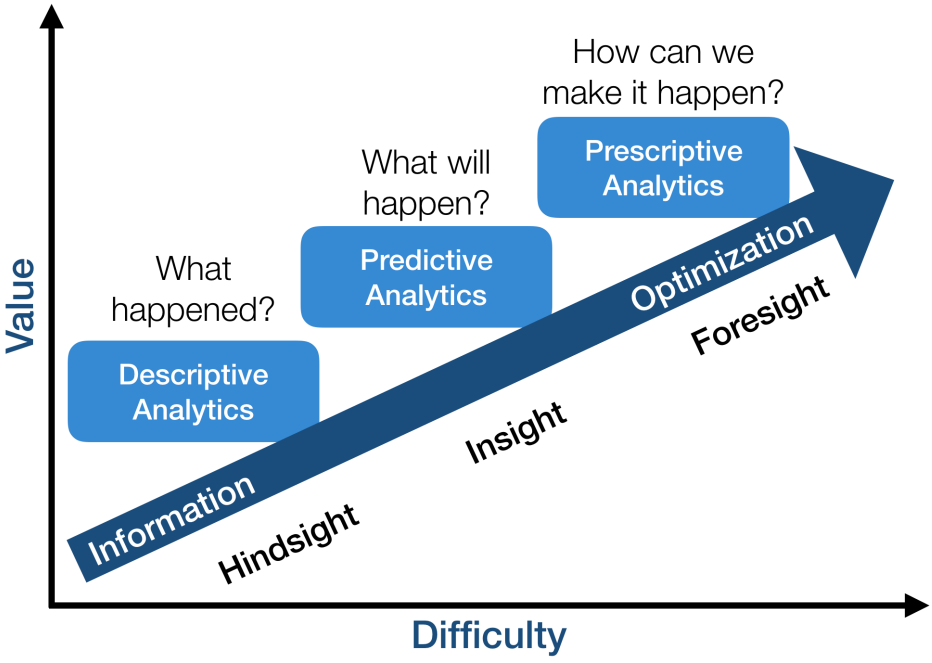


Figure 1.1: Descriptive, Predictive, and Prescriptive Analytics by Gartner

On the one hand, we will outline the **distinct types of analytics in data science**. For this purpose, Figure 1.1 depicts an infographic by Gartner illustrating Descriptive, Predictive, and Prescriptive Analytics along the two relevant dimensions, difficulty of implementation and potential for value creation. Furthermore, the infographic shows a directional arrow that emphasizes the overall goal of the analytics step, from purely describing information in hindsight up to optimization in foresight.

Descriptive Analytics focuses on the past, and attempts to generate information from existing data in order to explain what has happened. It is located at the bottom left of the infographic, which showcases a low difficulty of implementation and limited potential for creation of future value.

Predictive Analytics attempts to predict the future outcome based on the data at hand. It is located at the center of the infographic, with medium difficulty but also medium potential to create value.

Prescriptive Analytics aims to predict the future and also provides a detailed list of interventions that need to be followed in order to reach the optimal, most valuable future outcome out of all predicted possible futures. It is located at the top right of the infographic, with a high potential for value creation, but also a high difficulty of implementation.

On the other hand, we introduce the concept of the **Big Data Value Chain** in order to understand the process of value creation through Big Data Analytics. As illustrated in Figure 1.2, the Big Data Value Chain depicts a series of seven consecutive steps through which value can be extracted from Big Data [Miller & Mork 2013].

Each step entails one out of three major stages, namely **data discovery, data integration, or data exploitation**. The stage of data discovery contains three steps, collection and annotation, preparation, and organization of data. The stage of data integration only consists of a single step that brings the need for integration of datasets through a common representation of the data at hand into focus. Lastly, the stage of data exploitation contains the final three steps, which are data analytics, visualization, and decision making. In this final stage, after data has been collected, prepared, organized, and integrated, value can be created by utilizing the results of Big Data Analytics in order to positively influence the decision making process.

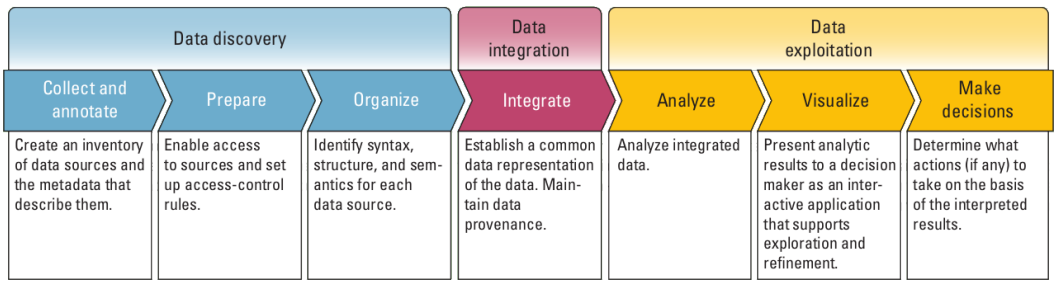


Figure 1.2: Big Data Value Chain [Miller & Mork 2013]

Based on the understanding of the three distinct types of data analytics and the value creation through the Big Data Value Chain, we introduce the **Big Social Data Analytics research framework** that is used by our research group at the Centre for Business Data Analytics. Similar to the three distinct types of data analytics from Gartner, our research framework implements a three-tiered approach based on **Descriptive, Predictive and Prescriptive Big Social Data Analytics**. This includes the creation of software tools such as Visual Analytics dashboards for Descriptive Analytics, forecasting models for Predictive Analytics, and recommender systems for Prescriptive Analytics. As illustrated in Figure 1.3, it covers all three stages of the **Big Data Value Chain**, in particular the data collection pipeline powered by the Social Data Analytics Tool [Hussain & Vatrapu 2014b], the merging of Big Social Data and enterprise data, and several analytical steps which are performed in order to produce relevant research findings.

The findings published by our research group based on this research framework can be seen as the results of a complex Big Data Value Chain, which is implemented from start to finish. Data collection is performed using the Social Data Analytics Tool, data integration is performed on-demand, and data exploitation is performed using the novel Social Set Analysis approach [Vatrapu *et al.* 2016], which will be detailed in the following chapters.

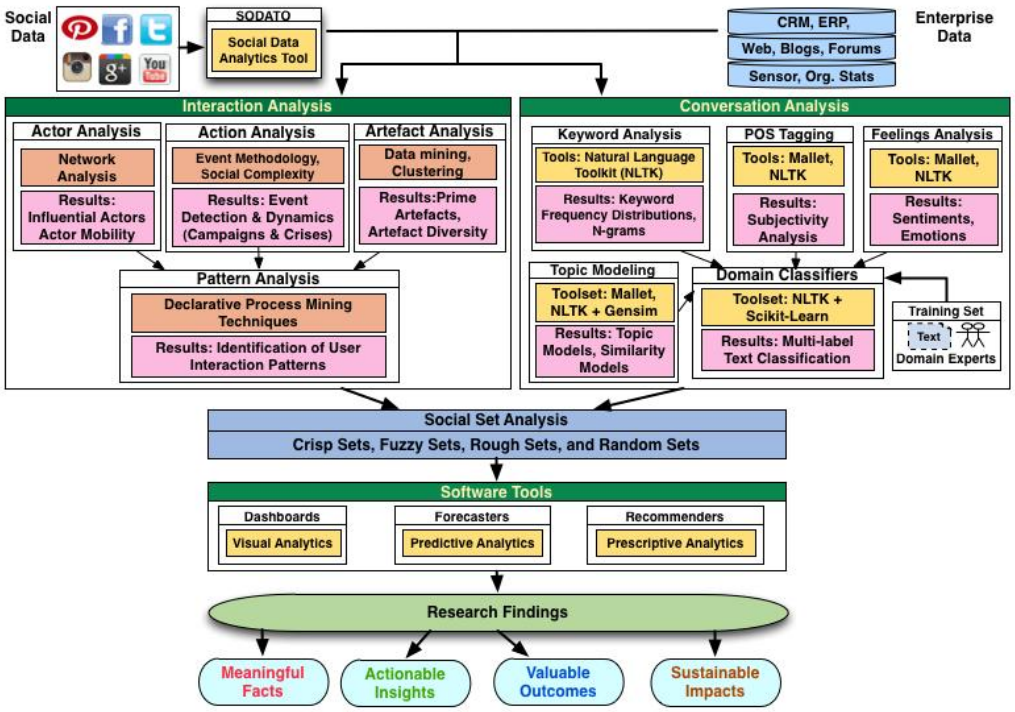


Figure 1.3: Big Social Data Analytics research framework with focus on Social Set Analysis [Vatrapu *et al.* 2016]

1.2 Research Problems

This PhD project is situated within the research framework presented in Figure 1.3, as part of the data exploitation process known from the Big Data Value Chain. There, it depicts a central component of the Social Set Analysis approach. It carries a particular thematical focus on **Visual Analytics of Big Social Data**, a subfield of Descriptive Analytics, that draws from the fields of vision science [Ware 2004], computer graphics [Munzner 2014] and human computer interaction [Jeffrey *et al.* 2010, Fisher *et al.* 2012], while retaining the exploratory data analysis focus of information visualization [Keim *et al.* 2008].

In order to outline the research problems addressed by this PhD project, first we **examine the state of the art in Visual Analytics of Big Social Data** through a systematic review of current academic literature. As part of a co-authored journal article on *Big Social Data Analytics: Past, Present, and Future* which is planned to be published soon [Vatrapu *et al.*], I have performed a systematic literature review of extant literature in Big Social Data Analytics with special focus on the state-of-the-art in Visual Analytics of Big Social Data. Articles were collected based on variations of the search terms "social media data", "visualization", and "analytics" from IEEE Xplore, ACM DL, Science Direct, and Scopus. They were further analyzed based on inclusion and exclusion criteria, such as only empirical studies which essentially focus on social media data and fit the 4V definition of Big Data. Furthermore, we rejected duplicates, non-English, and non-peer-reviewed articles.

For this review, 212 recent articles were selected from relevant scientific databases, and, after filtering based on the described criteria, 41 publications were reviewed in detail, as documented in Appendix A. The **results of the literature review** indicate a massive focus on Twitter-based datasets, with 27 (65.85%) of the analyzed articles using Twitter as their primary data source. The second most frequently used data source, Facebook, is only used in 10 (24.39%) of the surveyed articles. YouTube, the third largest data source, appears only in two articles (4.87%). Hence, Twitter and Facebook depict the two most commonly used sources of Big Social Data in state-of-the-art literature.

Taking a closer look at these two data sources, Figure 1.4 displays the number

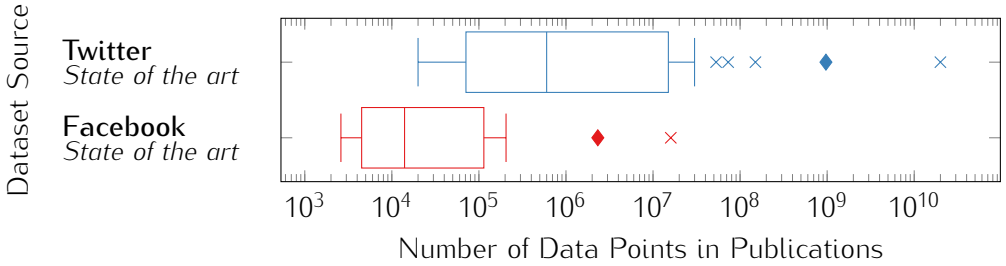


Figure 1.4: Comparison of dataset sizes from Twitter and Facebook from the reviewed literature on Visual Analytics of Big Social Data

of data points of each dataset that is used in the reviewed publications. Thereby, it highlights a **major imbalance between Twitter and Facebook** datasets. In the reviewed literature, both mean ($\approx 10^9$) and median ($\approx 10^6$) sizes of the Twitter datasets are significantly larger than the mean ($\approx 10^6$) and median ($\approx 10^4$) of the Facebook datasets.

This imbalance can also be highlighed by **comparing information density per unit of interaction between Big Social Data from Twitter and Facebook**. The average number of characters in each piece of user-generated content that is produced on both platforms can act as a simple proxy for information density per data point. Unfortunately, most publications don't state the total number of characters in their datasets, therefore this information had to be gathered from alternative sources, as illustrated in Table 1.1. In 2012, the average length of Facebook posts was 65 characters (n=24,009) and the average length of Facebook comments was 56 characters (n=75,381) [Guyot 2012]. In 2016, this PhD project observed the average length of Facebook posts as 173 characters (n=14,668) and the average length of Facebook comments as 44 (n=613,434) in a study on the US election based on Hillary Clinton and Donald Trump Facebook pages. In 2018, this PhD project observed the average length of Facebook posts as 213 characters (n=60,006) and the average length of Facebook comments as 101 characters (n=9,199,736) in a dataset consisting of 233 different Facebook pages in the area of entertainment, sports, and politics. In 2017, when Twitter had a maximum limit of 140 characters, the average length of Tweets was reported as 34 characters, while in 2018, when the Twitter limit was doubled to 280 characters, the average length of Tweets was slightly lower at 33 characters [Perez 2018].

The most recent data points emphasize that Facebook data on average has a greater number of characters per unit of interaction ($\approx$ 213 per post and $\approx$ 101 per comment) than Twitter data ($\approx$ 33 per Tweet). Thus, the **discrepancy in terms of information density** that is measured by comparing character counts in Facebook and Twitter datasets can be quantified at a factor of ≈ 3 to ≈ 6. Therefore, the major imbalance in dataset sizes between Twitter and Facebook studies that is shown in Figure 1.4 persists at several orders of magnitude, even when the factor of information density is taken into account.

Resulting from these two measurements, we can articulate a **research gap** with

Year	Average # of characters per data point			Source
	in FB posts	in FB comments	in Tweets	
2012	65	56		[Guyot 2012]
2016	173	44		US Election
2017			34	[Perez 2018]
2018	213	101		SoSeVi III
2018			33	[Perez 2018]

Table 1.1: Comparing average number of characters in Facebook and Twitter datasets

respect to the number of data points utilized in the state-of-the-art literature on Visual Analytics of Big Social Data. It becomes clear that the Facebook datasets used are several orders of magnitude smaller than the Twitter datasets. This observation implies that significant, contemporary difficulties exist for researchers in the field of Visual Analytics working with Facebook datasets.

Based on our understanding of the Big Data Value Chain, these difficulties can be caused by two potential problems that researchers face. On the one hand, it could be a **problem of data collection** from Facebook, and on the other hand, it could be a **problem of data exploitation**, once the data has been acquired.

First, we investigate the potential **problem of data collection** from Facebook. The question arises, whether large-scale Facebook datasets are significantly more difficult to acquire than Twitter datasets. If we look at relevant publications, we see that the particular challenge of Facebook data collection has already been overcome for several years. Furthermore, various publications from our research group such as on the Social Data Analytics Tool [Hussain & Vatrapu 2014b] give detailed description of large-scale data collection using the Facebook API. Therefore, data collection cannot be considered to be the predominant reason for this research gap.

Hence, the research gap is likely caused by an **ongoing problem of data exploitation** which negatively affects research on large-scale Facebook datasets. According to the Big Data Value Chain, the problem of data exploitation stems from a lack of analytical approaches that can reliably produce research findings from large-scale Facebook datasets. Moreover, it is also influenced by a lack of suitable visualizations that present analytic results. In the following, various arguments are presented to substantiate the nature of the problem.

Based on a survey of visualization techniques utilized in state-of-the-art literature, it can be observed that the **application of set visualization techniques to Big Social Data Analytics** depicts a unique and novel approach. Results of this survey are illustrated in Figure 1.5. On average, 3.07 different types of visualizations are used in each of the 41 reviewed publications. The most frequently utilized visualization techniques are maps (15x, 36.59%), line charts (13x, 31.71%), bar charts (12x, 29.27%), and timelines (12x, 29.27%). Furthermore, network graphs (11x, 26.83%), scatter plots (11x, 26.83%), and tables (9x, 21.95%) appear in a significant subset of the reviewed publications. We observe only infrequent use of heatmaps (6x, 14.63%), pie charts (4x, 9.76%), word clouds (4x, 9.76%), and radial plots (3x, 7.32%). No set-based visualization techniques have been found within the reviewed publications. Therefore, this PhD project is the **first to utilize set-based visualization techniques** for Visual Analytics of Big Social Data.

Major limitations in existing research on Big Social Data Analytics further underline the problem of data exploitation. We observe that computational methods, formal models, and software tools are largely limited to graph-theoretical approaches informed by relational sociology [Gross & Yellen 2005, Emirbayer 1997]. The most prominent graph-theoretical approach is **Social Network Analysis** [Borgatti *et al.* 2009, Boyd & Ellison 2007, Wasserman & Faust 1994, Tichy *et al.* 1979, Suthers 2017]. Previous work has established a lack of other unified modeling approaches to social data

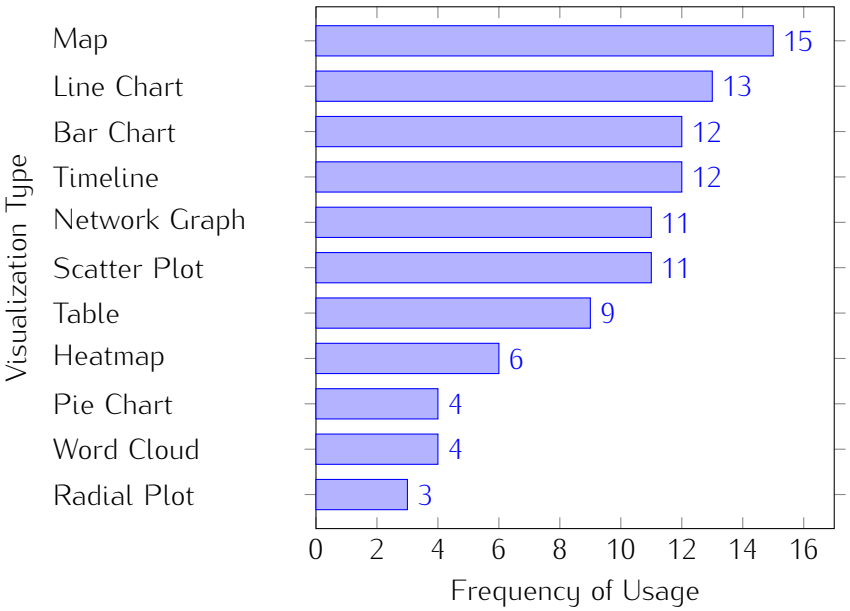


Figure 1.5: Utilization of visualization techniques within reviewed literature

apart from Social Network Analysis, which integrate the conceptual, formal, technological, analytical and empirical realms [Mukkamala *et al.* 2013]. Within the reviewed literature, Social Network Analysis is found to depict the main application domain for Visual Analytics of Big Social Data.

Data exploitation is further challenged when analyzing Big Social Data from platforms like Facebook, as such data consists of not only dyadic relations but also individual associations [Mukkamala *et al.* 2014]. On a high level, Social Network Analysis has two fundamental assumptions: First, that social reality is constituted by dyadic relations, and second, that interactions are determined by the structural position of individuals in social networks [Mizruchi 1994]. In order to generate insights from Big Social Data, these two assumptions are neither necessary nor sufficient [Vatrapu *et al.* 2014].

Subsequently, this PhD project aims to contribute to the advancement of the state of the art by applying a novel, set-based approach called **Social Set Analysis** to the unsolved problem of data exploitation, with particular relevance for Facebook datasets. Social Set Analysis is a set-based research approach situated in the domains of Data Science [Cleveland 2001, Loukides 2012, Ohsumi 2000] and Computational Social Science [Lazer *et al.* 2009] with practical applications to Big Social Data Analytics in organizations [Vatrapu 2013, Sterne 2010, Sponder 2012]. In recent years, it was devised and theoretically developed by our research group [Vatrapu *et al.* 2014, Vatrapu *et al.* 2016].

Due to its unique set-based approach, Social Set Analysis addresses **important theoretical and methodological limitations** in the emerging paradigm of Big Social Data Analytics, as existing approaches are mostly limited to graph-theoretical models

[Tufekci 2014]. In contrast to Social Network Analysis, which assumes homophily in its graph representation of the data, Social Set Analysis rather tries to capture the agentic mechanisms constituting homophily based on individuals within Big Social Data [Vatrapu *et al.* 2016].

Social Set Analysis is **closer to social reality and to social theory**, particularly to the concept of *Intersectionality* [Crenshaw 1990], a set-based formalization of social injustice which in recent years has become *"the primary analytic tool that feminist and anti-racist scholars deploy for theorizing identity and oppression"* [Nash 2008]. This concept is ideally modelled by Social Set Analysis and its set-based approach rather than with the graph-based approach of Social Network Analysis.

Moreover, Social Set Analysis provides a **fast and frugal community detection method** based on associations to entities, which can be ideas, identities, beliefs, causes, or other things. For many research questions, we are not interested into the structural characteristics of a problem, on which Social Network Analysis focuses, but into the analysis of set formations by entities of the same kind.

Furthermore, Social Network Analysis is **overly affected by incomplete data**, as fundamental network metrics can substantially change after the addition or removal of individual actors [Wei *et al.* 2016]. Meanwhile, results from a set-based approach are significantly less affected by this problem.

Handling of networks in Social Network Analysis, most importantly the **processing and clustering of large graphs**, depicts a challenging computational problem. It returns a wide array of possible clustering results depending on the chosen parameters, with parameter values often requiring further interpretation. In contrast, Social Set Analysis provides a **simple-to-understand, straightforward-to-execute methodology** based on the mathematics of set theory. Social Set Analysis computations are mainly limited by available working memory due to large set cardinalities, and not limited by CPU or GPU performance, as it is the case with computation and visualization of large-scale graphs in Social Network Analysis. They can be performed in a divide-and-conquer approach based on partitioned subsets of data, therefore lending themselves to a vast array of parallelization and caching strategies without negatively affecting the overall validity of computation results.

Consequently, this PhD project aims to significantly reduce the outlined research gap. On the one hand, the mean ($\approx 10^8$) and median ($\approx 10^8$) number of data points in my research publications is considerably larger than for the Facebook datasets in the

Datasets			Number of Data Points	
Source	Type	n	Mean	Median
Twitter	State of the art	27	$\approx 10^9$	$\approx 10^6$
Facebook	State of the art	10	$\approx 10^6$	$\approx 10^4$
Facebook	**This PhD project**	**6**	$\approx \mathbf{10^8}$	$\approx \mathbf{10^8}$

Table 1.2: Comparing the number of data points in state-of-the-art research on Visual Analytics of Big Social Data vs. this PhD project

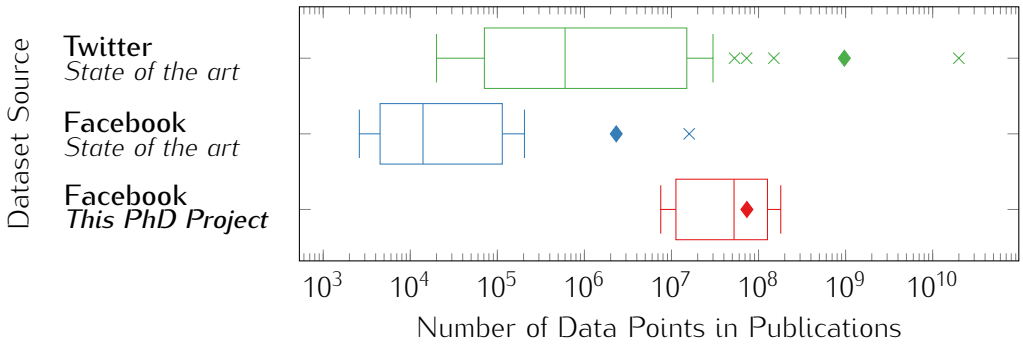


Figure 1.6: Comparison of dataset sizes from Twitter and Facebook from the reviewed literature on Visual Analytics of Big Social Data vs. this PhD project

reviewed literature, as detailed in Table 1.2. This is possible in part due to previous contributions by our research group to data collection, most importantly through the Social Data Analytics Tool [Hussain *et al.* 2014]. Overall, the **number of data points is greatly increased** by 100× in relation to the mean, and by 10,000× in relation to the median of the Facebook dataset sizes in state-of-the-art research. The contribution of my PhD project towards this research gap is illustrated in Figure 1.6. On the other hand, this PhD project supplements the established Social Network Analysis approach with Social Set Analysis, which **amplifies the generation of insights from Big Social Data** in particular from Facebook datasets. Thus, a practical Big Social Data Analytics approach inspired by Social Set Analysis needs to be researched and demonstrated. The presentation and evaluation of the Social Set Visualizer (SoSeVi) software tool in this thesis significantly contributes to the resolution of this research problem.

Through this, my PhD projects helps to advance the state of the art in research on Visual Analytics of Big Social Data in terms of large-scale Facebook datasets. This **advancement of the state of the art** is only possible because it aims to solve two hard research problems:

First, it solves the prevailing research problem of **generating meaningful insights from Big Social Data** through implementation and evaluation of the Social Set Visualizer software tool. Resolution of this problem is possible through utilization of a novel set-based approach to Big Social Data Analytics called Social Set Analysis that was devised by our research group. This resolves an important problem for many researchers not only due to the sheer volumes of data involved, but also due to the limitations of current methodologies, as emphasized by the literature review.

Secondly, it solves the **difficult technical challenge of interactively visualizing large-scale sets and set intersections** from the field of Visual Analytics. This challenge arose both during design and implementation of the Social Set Visualizer software tool, as large-scale sets calculated from Big Social Data need to be accurately analyzed and visualized. It is resolved through the application of innovative set visualization techniques to Big Social Data Analytics, which in turn allows the Social

Set Visualizer software tool to utilize the novel Social Set Analysis approach to its full potential. Interactivity of the software tool speeds up the important processes of data analysis and insight generation, which in turn facilitates an iterative approach to answering research questions.

The outlined research gap displays **both a research problem and a technical problem**. On the one hand, it is a research problem because larger datasets are needed to present more general findings, to limit the impact of biases, and to effectively compare different cohorts of data. On the other hand, the research gap is a technical problem insofar that many researchers are unable to work with datasets beyond a certain size, as evidenced by the literature review, and that the creation of software tools, even moreso interactive software tools, remains a difficult challenge that requires significant experience in software engineering.

Hence, significant difficulties with regard to computation and visualization need to be overcome during design and development of the Social Set Visualizer software tool that is presented in this thesis.

1.3 Research Questions

My PhD projects takes on the challenge of resolving the research problems elaborated in the previous section. With this thesis, I aim to contribute to current research via design, development and evaluation of the Social Set Visualizer, which is based on a revised theoretical model of Big Social Data, namely the Social Interaction Model. The Social Set Visualizer is a cutting-edge Visual Analytics software tool for Big Social Data that depicts a tailormade IT artifact for the novel Social Set Analysis approach. It has been developed based on several years of research towards tools and methodologies for Social Set Analysis as well as a variety of iterations following the Action Design Research methodology.

Given this context and the previously derived research problems, this thesis has the objective to answer the following two research questions:

> RQ1: *How and in what way can the novel Social Set Analysis approach to Big Social Data Analytics be modeled into an interactive Visual Analytics software tool that can be utilized for generating meaningful insights from Big Social Data?*
>
> RQ2: *What are software design requirements for a Visual Analytics software tool that interactively visualizes large-scale sets and set intersections with multiple users and large amounts of data?*

1.4 List of Publications

This dissertation consists of multiple research publications that have been peer-reviewed and presented to an international academic audience through journals and

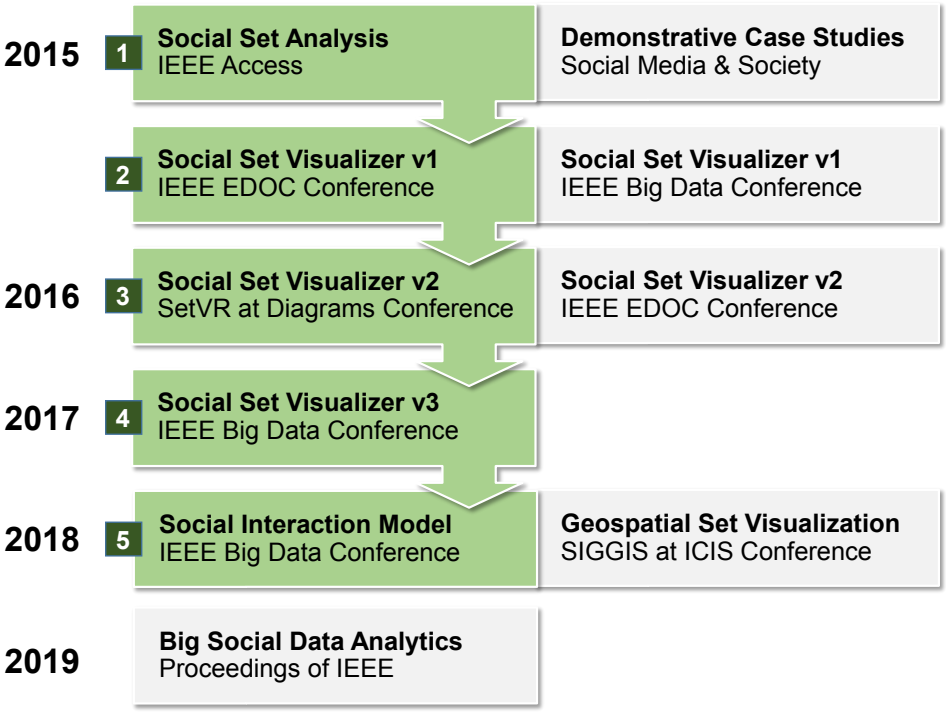


Figure 1.7: Overview of relevant peer-reviewed publications on the Social Set Visualizer with highlight on five publications chosen for this thesis

conferences. The presented work stems from more than three years of systematic, iterative research which focused on the theories and applications of Social Set Analysis in Big Social Data Analytics. Throughout my PhD studies, I have co-authored a total of **13 peer-reviewed publications in the field of Big Social Data Analytics**. Out of these publications, seven are first-authored.

Five of these 13 publications have been chosen to be included in this thesis and to be given a special focus due to their relevance to the Social Set Visualizer. Each of these five chosen publications contributes to at least one aspect of the Social Set Visualizer, either towards the design, development or evaluation. This includes a single-authored publication which extends and refines the conceptual model of social data based on various learnings from the research presented in this thesis.

The five peer-reviewed focus publications that have been selected for this thesis are highlighted in the publication overview with regard to the 10 publications on the Social Set Visualizer (see Figure 1.7). They are accompanied by further publications relevant to the development of the Social Set Visualizer, which are not explicitly included in this thesis. Three further student papers are also not illustrated in this figure. In the following, a short overview on each of the five included publications will be given.

Publication 1: Journal article introducing Social Set Analysis approach

Ravi Vatrapu, Raghava Rao Mukkamala, Abid Hussain and **Benjamin Flesch**. *Social Set Analysis: A Set Theoretical Approach to Big Data Analytics*. IEEE Access: Special Section on Theoretical Foundations for Big Data Applications: Challenges and Opportunities, vol. 4, pages 2542–2571, 2016

This journal article introduces the novel Social Set Analysis approach to a larger academic audience. It develops the theoretical foundation of a set-based approach to Big Data Analytics. Furthermore, it outlines future work with regard to the implementation of software tools that incorporate the Social Set Analysis approach in order to improve the generation of insights from Big Social Data. It was published in a special edition of IEEE Access, which is dedicated to the advancement of the theoretical foundations of Big Data, with particular focus on challenges and opportunities.

Publication 2: Presentation of the Social Set Visualizer

Benjamin Flesch, Abid Hussain and Ravi Vatrapu. *Social Set Visualizer: Demonstration of Methodology and Software*. In 2015 IEEE 19th International Enterprise Distributed Object Computing Workshop, pages 148–151, Sept 2015

This conference paper presents the first version of the Social Set Visualizer software tool at IEEE EDOC. It showcases the set-based approach to Big Social Data Analytics, and presents the web-based Visual Analytics dashboard. Furthermore, it generates insights on social media reactions to international clothing retailers in the wake of the 2013 Bangladesh factory disasters, utilizing a Facebook dataset of 180M data points.

Publication 3: Social Set Visualizer with UpSet-style Visualizations

Benjamin Flesch, Raghava Rao Mukkamala, Abid Hussain and Ravi Vatrapu. *Social Set Visualizer (SoSeVi) II: Interactive Social Set Analysis of Big Data.* In SetVR@ Diagrams, pages 19–28, 2016

This conference paper presents the second version of the Social Set Visualizer at the Set Visualization and Reasoning (SetVR) workshop colocated with Diagrams conference in Philadelphia, PA. In this paper, the Social Set Visualizer software tool presents novel, scalable visualizations of large-scale sets and set intersections inspired by the UpSet approach [Lex *et al.* 2014] and applies this approach to Social Set Analysis of Big Social Data. Furthermore, it showcases a set-based visualization of migration patterns in social media. The audience of this highly specialized workshop consisted of set visualization experts, who were very curious about the utilization of cutting-edge set-based visualization techniques and their application to the problem of Big Social Data Analytics.

Publication 4: Final Iteration on the Social Set Visualizer

Benjamin Flesch, Ravi Vatrapu and Raghava Rao Mukkamala. *A Big Social Media Data Study of the 2017 German Federal Election Based on Social Set Analysis of Political Party Facebook Pages with SoSeVi*. In Big Data (Big Data), 2017 IEEE International Conference on, pages 2720–2729. IEEE, 2017

This conference paper presents the third version of the Social Set Visualizer, which includes a built-in data collection process, and thereby represents a complete implementation of the entire Big Data Value Chain. It contains a comprehensive case study on the 2017 German parliamentary election. Deep insights are generated through use of the Social Set Visualizer software tool on a large-scale Facebook dataset with more than 15M data points. Furthermore, it includes a custom query language which can be interactively used in order to create visualizations of large-scale set intersections from Big Social Data.

Publication 5: Unified Social Interaction Model for Big Social Data

Benjamin Flesch. *Social Interaction Model*. In Big Data (Big Data), 2018 IEEE International Conference on. IEEE, 2018

This single-authored conference paper presents the Social Interaction Model, a unified model of Big Social Data, which extends and supplements the existing model of social data [Mukkamala *et al.* 2013] that was previously developed by our research group. Based on the experiences gathered by utilizing the Social Set Analysis approach to Big Social Data Analytics over the course of my PhD project, it proposes a framework based on *Actions* and *Reactions* in social media. From these *Interactions*, various types of textual and non-textual *Artifacts* are created. Thereby, it radically simplifies the existing models and is in line with analytical approaches that utilize dimensions of time and space for a set-based analysis, such as the Social Set Analysis approach.

Further Publications

Furthermore, eight peer-reviewed publications have not been included in this thesis. Even though these eight excluded publications provide further case studies of the Social Set Visualizer, the five chosen focus publications provide a good picture on the contributions of this PhD project.

- **Benjamin Flesch**, Ravi Vatrapu, Raghava Rao Mukkamala and Abid Hussain. *Social Set Visualizer: A Set Theoretical Approach to Big Social Data Analytics of Real-world Events*. In Big Data (Big Data), 2015 IEEE International Conference on, pages 2418–2427. IEEE, 2015
- Ravi Vatrapu, Abid Hussain, Niels Buus Lassen, Raghava Rao Mukkamala, **Benjamin Flesch** and Rene Madsen. *Social Set Analysis: Four Demonstrative Case Studies*. In Proceedings of the 2015 International Conference on Social Media & Society, page 3. ACM, 2015

- **Benjamin Flesch** and Ravi Vatrapu. *Social Set Visualizer (SoSeVi) II: Interactive Computational Set Analysis of Big Social Data*. In Enterprise Distributed Object Computing Workshop (EDOCW), 2016 IEEE 20th International, pages 1–4. IEEE, 2016

- Linda Camilla Boldt, Vinothan Vinayagamoorthy, Florian Winder, Melanie Schnittger, Mats Ekran, Raghava Rao Mukkamala, Niels Buus Lassen, **Benjamin Flesch**, Abid Hussain and Ravi Vatrapu. *Forecasting Nike's sales using Facebook data*. In Big Data (Big Data), 2016 IEEE International Conference on, pages 2447–2456. IEEE, 2016

- Anna Hennig, Anne-Sofie Åmodt, Henrik Hernes, Helene Nygårdsmoen, Peter Arenfeldt Larsen, Raghava Rao Mukkamala, **Benjamin Flesch**, Abid Hussain and Ravi Vatrapu. *Big Social Data Analytics of Changes in Consumer Behaviour and Opinion of a TV Broadcaster*. In Big Data (Big Data), 2016 IEEE International Conference on, pages 3839–3848. IEEE, 2016

- **Benjamin Flesch**, Ravi Vatrapu, Raghava Rao Mukkamala and René Madsen. *Real-time Geospatial Visualization of Crowd Trajectory at Roskilde Festival 2018*. In ICIS 2018 Special Interest Group on Geographic Information Systems (SIGGIS) Pre-Conference Workshop Proceedings. 1., SIGGIS '18. ACM, 2018

- Tor-Morten Groenli, **Benjamin Flesch**, Raghava Rao Mukkamala and Ravi Vatrapu. *Internet of Things Big Data Analytics: The Case of Noise Level Measurements at the Roskilde Music Festival*. In Big Data (Big Data), 2018 IEEE International Conference on. IEEE, 2018

- Ravi Vatrapu, Hannu Kärkkäinen, Raghava Rao Mukkamala, Karan Menon, Jukka Huhtamäki, Jari Jussila, **Benjamin Flesch** and Niels Buus Lassen. *Big Social Data Analytics: Past, Present, and Future*. Unpublished Manuscript (Work in progress)

1.5 Thesis Outline

This section briefly outlines the chapters of the dissertation, while highlighting the contribution of individual publications to each chapter.

Chapter 2: Research Methodology

The second chapter provides details on the research methodology of this dissertation. It introduces Action Design Research and the analytical techniques of Social Set Analysis and Event Study Methodology. Furthermore, it presents the conceptual models of Big Social Data, namely the Social Data Model and the Social Interaction Model. Lastly, it specifies the data collection in this PhD project and lists the utilized datasets.

Chapter 3: Design

The third chapter of this dissertation details the design of the Social Set Visualizer. It introduces the target audience, design goals, and design objectives of the IT artifact. Subsequently, the three user interfaces of the software tool are presented, namely the browser-based Visual Analytics dashboard, the visual query builder, and the textual query language. Furthermore, the state of the art in the visualization of sets is highlighted, explaining Euler and Venn diagrams, the EulerAPE approach, and linear diagrams. In addition, recent approaches to set visualization, namely linear diagrams, UpSet and UpSetR, are introduced. The implemented visualizations in the Social Set Visualizer are presented, in particular the UpSet- and UpSetR-styled set visualizations and the approach of "exploded" Venn diagrams.

Chapter 4: Development

The fourth chapter concerns the development of the Social Set Visualizer. It outlines development objectives and technological foundations in terms of data storage and visualizations. Additionally, the software architecture for frontend and backend is detailed. Then, it presents the three iterations on the Social Set Visualizer. Lastly, it describes the deployment of the software tool.

Chapter 5: Evaluation

In the fifth chapter, the Social Set Visualizer is evaluated through seven case studies. Four case studies utilize the software tool for descriptive analytics on the topics of corporate social responsibility, sports broadcasting, music festivals, and emission scandals. Furthermore, three case studies utilize it for predictive analytics in case studies on sales forecasting, concert audience prediction, and election prediction.

Chapter 6: Discussion

In the penultimate chapter, the work presented in this thesis is discussed, with a view on its implications and limitations. A particular reflection is made on research methodology, visualization of sets, the presented IT artifact, the theoretical data model, and the domain-specific query language.

Chapter 7: Conclusions and Future Work

The final chapter summarizes the findings of this PhD projects and concludes this thesis. It highlights the theoretical and practical contributions and outlines potential future work with regard to Big Data Analytics and in particular the Social Set Visualizer.

Chapter 2

Research Methodology

This chapter presents sets methodological foundations of this PhD project. The underlying methodologies used in the various publications of this PhD project are summarized in the following in order to provide an introduction to the reader. Starting from the overarching methodology of Action Design Research, the utility of its iterative, artifact-based approach for this PhD project is showcased. Consequently, the two analytical techniques which are utilized for insight generation in the Social Set Visualizer software tool are presented, namely the Social Set Analysis approach and the Event Study Methodology. Moreover, the theoretical models of Big Social Data are introduced, which have been developed by our research group and serve as a basis for the papers in this thesis, alongside with related work on the topic of modelling socio-technical interactions. Furthermore, this dissertation practically applies and theoretically extends the existing *Social Data Model*. Hence, both the *Social Data Model* and its proposed enhancement, the *Social Interaction Model*, are specified in this chapter. Subsequently, the collection of Big Social Data from Facebook is outlined and a list of datasets is given. Lastly, the analytical processes utilized in this dissertation are detailed and examples for set-based approaches to common analytical questions are outlined. This depicts the foundation of all analytical work presented in this dissertation.

2.1 Action Design Research

This thesis implements the Action Design Research methodology, developed by [Sein *et al.* 2011]. Action Design Research is an information systems research framework which is grounded in Design Science methodology [Vaishnavi & Kuechler 2004, Hevner 2007, Collins *et al.* 2004]. It extends the concept of Design Science with the goal to improve organizational capability through development of technological innovations that are fed back into the organizational information systems. Central element of Action Design Research is the creation and refinement of an IT artifact, in consideration of both the technological and the organizational context.

The conceptual foundation of Action Design Research consists of four stages, namely problem formulation; building, intervention, and evaluation; reflection and learning; and the formalization of learning. These four stages are illustrated in Figure 2.1, originally published in [Sein *et al.* 2011].

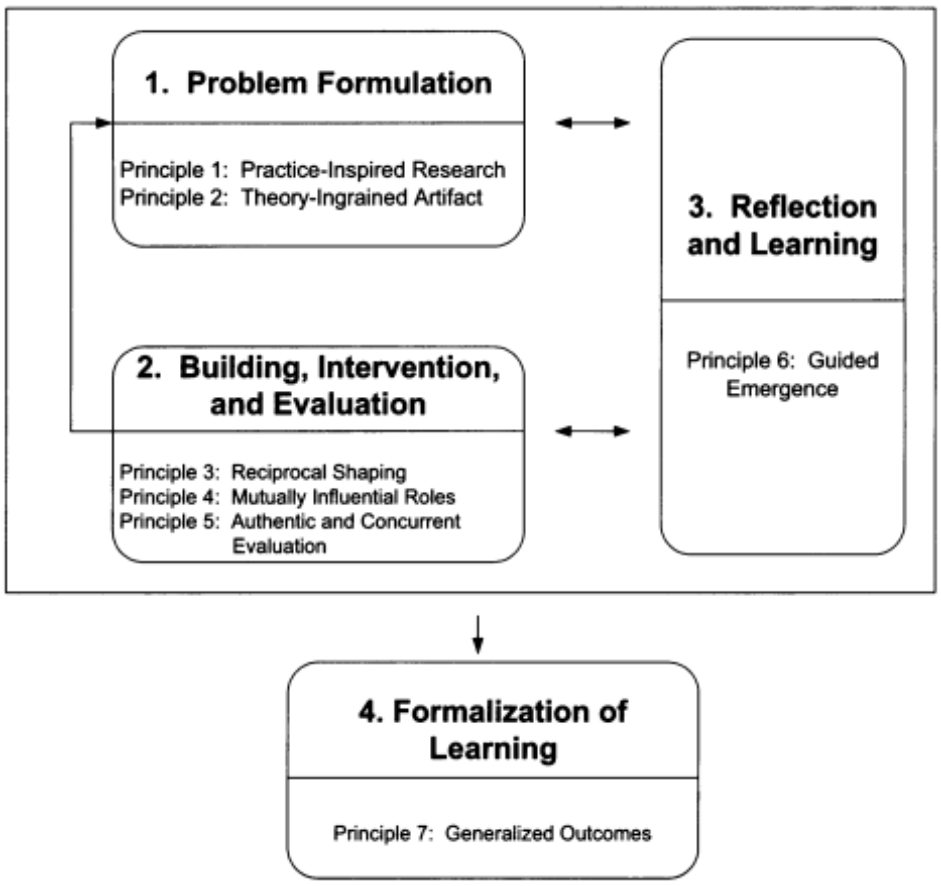


Figure 2.1: Stages and Principles in Action Design Research [Sein *et al.* 2011]

Stage 1: Problem Formulation

First, the initial problem is clearly formulated. At this stage the central focus of the researcher lays on principles of practice-inspired research and theory-ingrained artifacts. On the one hand, real-world problems are used to identify research opportunities where organizational value can be realized through design of an IT artifact. On the other hand, the creation of IT artifacts is built on a strong foundation of state-of-the-art theory.

Stage 2: Building, Intervention and Evaluation

Second, the building, intervention and evaluation (BIE) stage is performed in context of the Action Design Research framework. The BIE stage consists of three core principles, namely reciprocal shaping, mutually influential roles, and authentic and concurrent evaluation. These principles intend to catalyze an iterative process at the intersection of the IT artifact and the organizational environment [Sein *et al.* 2011].

The BIE comes in two specialized versions, one **IT-dominant BIE** and one **organization-dominant BIE**. For this thesis, the IT-dominant BIE is chosen, as this approach suits

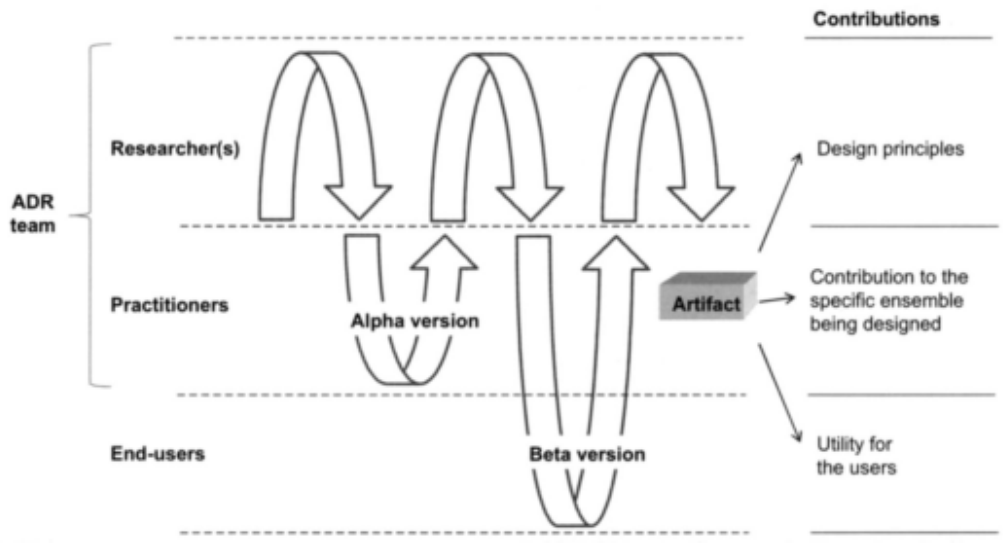


Figure 2.2: Generic Schema for IT-Dominant BIE [Sein *et al.* 2011]

Action Design Research efforts that emphasize creating an innovative technological design at the outset [Sein *et al.* 2011]. The IT-dominant BIE is illustrated in Figure 2.2. Its main stakeholders are researchers, practicioners and end-users.

Stage 3: Reflection and Learning

Third stage of Action Design Research is designed to reflect on and to learn from the two previous stages. Intervention results are analyzed with the obective of planning next steps and neccessary consequences. The principle of guided emergence combines the previously introduced principles of creating a theory-ingrained artifact, reciprocal shaping, mutually influential roles, and authentic and concurrent evaluation.

In this process, the initial design is presented by the researchers based on a certain working theory. It is then shaped by organizational use, and reflected on in a redesign [Henfridsson 2011].

Stage 4: Formalization of Learning

Within the last stage of Action Design Research, learnings from the previous stages are formalized. These learnings enable the researchers to reach generalized outcomes which can be transferred and applied to other problem areas.

Suitability of Action Design Research for this PhD Project

Action Design Research methodology has been selected for this PhD project due to its particular suitability from numerous angles. First, similar to this dissertation,

Action Design Research's main objective is to substantially improve organizational capabilities in the context of Big Social Data Analytics.

Second, the necessary timeline for integration of iterations and learnings within Action Design Rearch methodology is met. The multi-year duration of this PhD project lends itself to the formalization of the many iterations during creation of the IT artifact and resulting learnings within the Action Design Research framework. In accordance with this, the Social Set Visualizer software tool presented in this thesis depicts an IT artifact that incorporates novel design principles for set-based Visual Analytics. These novel principles have been iteratively designed, developed, and evaluated through various case studies during the course of this PhD project.

Furthermore, the research presented in this thesis provides interfaces with several outside stakeholders. Consequently, utility is created through the use of Big Social Data Analytics for the generation of insights.

The analysis of contemporary theory behind Social Set Analysis, the Social Data Model, and my extended and simplified theory, utimately resulting in the proposal of the Social Interaction Model, is in line with core principles of Action Design Research. Th chain of publications presented in this dissertation lends itself to the idea of a programmatic research stream [Nunamaker *et al.* 2017], a concept that has recently received momentum with design science scholars. Hence, my PhD project follows Action Design Research through development of new ideas and improvement of existing theories across several publications, which makes it well-aligned with current research in the academic design science community.

2.2 Analytical Techniques

In this section, the analytical foundations for generation of insights using the Social Set Visualizer are provided. It consists of Social Set Analysis methodology along the two dimensions of space and time, and Event Study Methodology.

2.2.1 Social Set Analysis

Social Set Analysis as employed in this PhD project is concerned with the mobility of social media actors across the two dimensions of time and space. The concept of Social Set Analysis is detailed in **Publication I** [Vatrapu *et al.* 2016] of this dissertation.

For mobility across time, we create a set of actors that interacted with a certain Facebook wall before, during and after a real-world event of interest. Then, set intersections between the three sets depicting the before, during and after time periods are calculated. Similarly, for mobility across space, set inclusion and exclusion is performed based on the different Facebook walls with which social media actors have interacted.

This set-based methodology enables us to uncover and quantify the interactional dynamics in Big Social Data. If set comparisons across time and space are combined with other filters, results correspond to marketing segments such as brand loyalists, brand advocates, brand critics, and social activists.

Mobility of Actors across Time

As part of Social Set Analysis, we have considered three different time frames for an event: before, during and after. This corresponds to pre-event, event and post-event timelines of the Event Study Methodology. For an event, sets containing unique actors who performed interactions during, before and after are computed. Respectively, the during actors set contains the actors who have either posted or commented or liked an artifact in the pre-event time period. With regard to before and after actors sets, the unique actor sets can be computed easily by adapting the reference timestamps to before and after the event period. Finally, intersections between actor sets are computed using standard set operations. As an example, it is possible to examine the set of unique actors who have performed actions only during the event period (neither before nor after) in order to identify potential patterns and demographic attributes that are unique to the event time period.

Mobility of Actors across Space

In Social Set Analysis, mobility across space corresponds to a notion of actors interacting with different Facebook walls. Given a set of Facebook walls, actors mobility across space can be computed through the set of actors who have interacted with all walls. Mobility across space is useful for analytical purposes in domains ranging from brand loyalty (actors who have visited only one wall) to social activism (actors who have visited many walls and interacted respectively, e.g. to express their protest).

Mobility of Actors across Time and Space

By combining the analysis of mobility across time and space, we can compute sets of actors that have interacted within a specific time period, e.g. during an event, and that also have interacted with certain other Facebook walls. By calculating the intersection along both dimensions, it is possible to add value beyond individual insight generation per dimension.

2.2.2 Event Study Methodology

Event studies depict a finance methodology to assess an impact on corporate wealth (e.g. stock prices) caused by events such as restructuring, leadership change, mergers, and acquisitions [Bromiley *et al.* 1988,McWilliams & Siegel 1997,MacKinlay 1997]. It has been a powerful tool since the late 1960s and was used exclusively in the area of finance, in particular to examine stock price performance and the dissemination of new information [Binder 1998]. However, concepts of Event Study Methodology are applicable to other research problems. Thus, Event Study Methodology is used in the context of Big Social Data Analytics.

While there is no unique structure for Event Study Methodology, at a higher level of abstraction, it includes identifying three main time periods of an event of interest. First, identifying the period over which the event is active (event window), second,

identifying the estimation period for the event (pre-event or estimation window), and third, identifying the post-event window [MacKinlay 1997]. When using the Social Set Analysis approach to analyze a real-world event, we apply Event Study Methodology to identify the three important time periods of user interactions on social media platforms, namely the pre-event window (*before*), the event window (*during*), and the post-event window (*after*).

2.3 Theoretical Models of Big Social Data

In this section, the theoretical models of Big Social Data underlying the analytical approach of the Social Set Visualizer software tool are introduced. During the course of this PhD project, the pre-existing theoretical data model for set-based Big Social Data Analytics, the Social Data Model, has been refined, simplified, and adapted based on the learnings from the development of the Social Set Visualizer.

Therefore, this section first introduces the theoretical model of socio-technical interactions from 2010 [Suthers *et al.* 2010], the classical Social Data Model from 2013 [Mukkamala *et al.* 2013] and its updated version from 2016 [Vatrapu *et al.* 2016]. Along this PhD project, the Social Data model is further extended which led to the development of the Social Interaction Model. This novel conceptual model for Big Social Data has been published in **Publication V** [**Flesch** 2018] of this thesis.

2.3.1 Theoretical Model of Socio-technical Interactions (2010)

Prior research on theoretical models of socio-technical interactions was introduced in [Suthers *et al.* 2010] and expanded in [Suthers & Rosen 2011], with special focus on distributed learning and knowledge creation in context of socio-technical networks. In this research, socio-technical interactions are also observed as being "distributed across actors, space and time" [Suthers *et al.* 2010], even though explicit traces of interactions are not always available for research purposes. Furthermore, the researchers introduce the concept of "uptake", where actors take contents produced by someone else and utilize it in their own communications, all without explicitly referencing or even knowing about the original author of the content.

2.3.2 Social Data Model (2013)

In 2013, the Social Data Model was first proposed as a theoretical model for Big Social Data [Mukkamala *et al.* 2013, Mukkamala *et al.* 2014]. Its concept of social data draws from the **theory of socio-technical interactions** [Vatrapu 2010], which describes the interactions of actors in social media. As the theoretical model does not distinguish between users and actors, the term actor is utilized for the remainder of this thesis.

According to the model, *Social Data* consists of two main categories, namely *Social Graph* and *Social Text*, as illustrated in Figure 2.3.

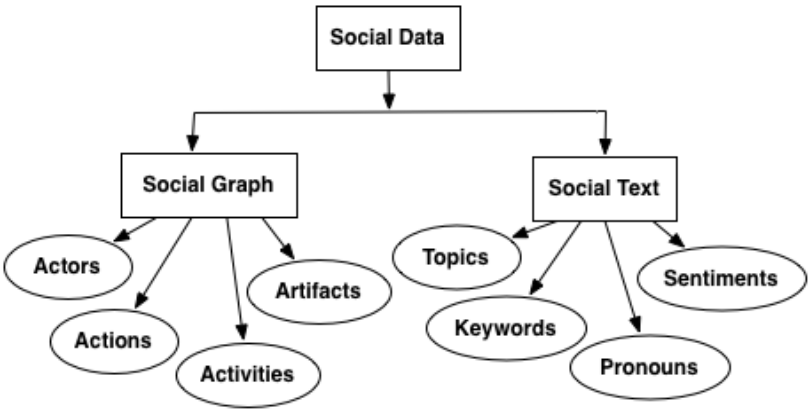


Figure 2.3: Social Data Model [Mukkamala *et al.* 2013]

Social Graph maps on to the first aspect of socio-technical interactions that involve perception and appropriation of affordances, e.g. the question of which actors utilize which technological features in order to interact with which other actors in the social media systems. It focuses on identifying the *Actors* involved, the *Actions* they take, the *Activities* they undertake, and the *Artifacts* they create and interact with. With respect to the terms *Action* and *Activity*, an *Action*, e.g. a post, comment, or like, is an atomic event done by an *Actor* on an *Artifact*, whereas an *Activity*, e.g. a promotion or campaign, can spread across many *Actions*, *Artifacts* and *Actors*. Hence, the *Social Graph* consists of the **structure of the relationships** emerging from the appropriation of social media affordances such as posting, linking, tagging, sharing and liking.

Social Text maps on to the second aspect of socio-technical interactions that constitute the structures and functions and technological intersubjectivity, e.g. which users, or actors, are trying to communicate to each other and how they are trying to influence each other through language. The *Social Text* consists of **communicative and linguistic aspects** of the social media interaction such as the *Topics* discussed, *Keywords* mentioned, *Pronouns* used and *Sentiments* expressed.

According to the authors, the purpose of the Social Data Model is to **prepare the application of computational tools and techniques** to Big Social Data [Mukkamala *et al.* 2014]. The Social Data Analytics Tool, which is introduced in section 2.4.2, depicts the first software implementation of the Social Data Model [Hussain & Vatrapu 2014a].

2.3.3 Updated Version of the Social Data Model (2016)

In 2016, the Social Data Model was revised based on several years of in-depth research on Big Social Data Analytics. This update to the model was included in **Publication I** [Vatrapu *et al.* 2016] in this dissertation which refined and formalized

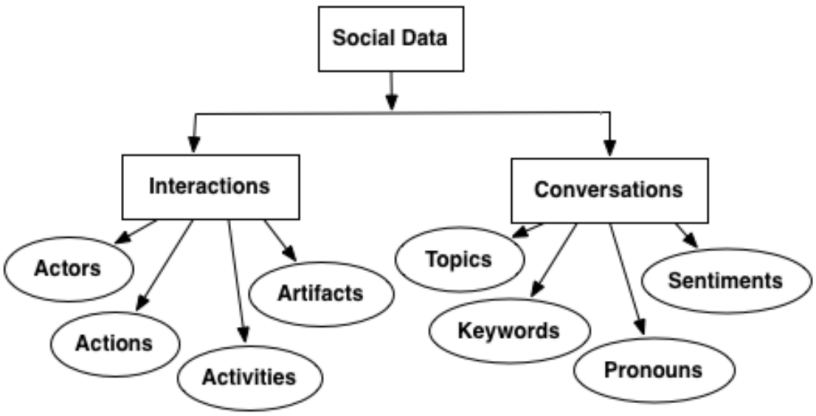


Figure 2.4: Updated version of the Social Data Model [Vatrapu *et al.* 2016]

the Social Set Analysis approach. The updated version of the Social Data Model is illustrated in Figure 2.4. Key changes entail the renaming of the concept of *Social Graph* to *Interactions*, and the concept of *Social Text* to *Conversations*. Except from this **change of names**, no further modifications were applied to the model. Therefore, the theoretical foundations of the original Social Data Model persist.

2.3.4 Social Interaction Model (2018)

In 2018, I presented an addition to the existing theoretical model of Big Social Data, developing the Social Interaction Model. The Social Interaction Model extends the existing Social Data Model through proposing of a radically simplified concept, which is **formally grounded in set theory and relational algebra**. The Social Interaction Model is included as **Publication V** [**Flesch** 2018] of this dissertation and contributes to the state-of-the-art theory of Big Social Data.

The conceptual foundation of the Social Interaction Model is illustrated in Figure 2.5. The Social Interaction Model distinguishes between three major components of social data: ***Actors, Interactions* and *Artifacts***. These core components of the Social Interaction Model show strong commonalities with the Social Data Model. The Social Interaction Model incorporates several foundational concepts from the Social Data Model, namely *Social Data*, *Interactions*, *Actors*, *Actions*, *Artifacts*, *Conversations*, *Topics*, *Keywords*, *Pronouns*, and *Sentiments*. At the same time the Social Interaction Model extends it with novel concepts of *Reactions*, *Social Video*, *Social Images*, and *Social Text*, which are displayed with thick borders in Figure 2.5. It reorganizes the two pillars of *Interactions* and *Conversations* proposed by the previous Social Data Model into a streamlined top-down flow, where *Social Text*, which is renamed from the former concept of *Conversations*, is extended by *Social Video* and *Social Images*. The Social Interaction Model expresses that all three types of *Artifacts*, namely *Social Video*, *Social Images*, and *Social Text*, can be analyzed for

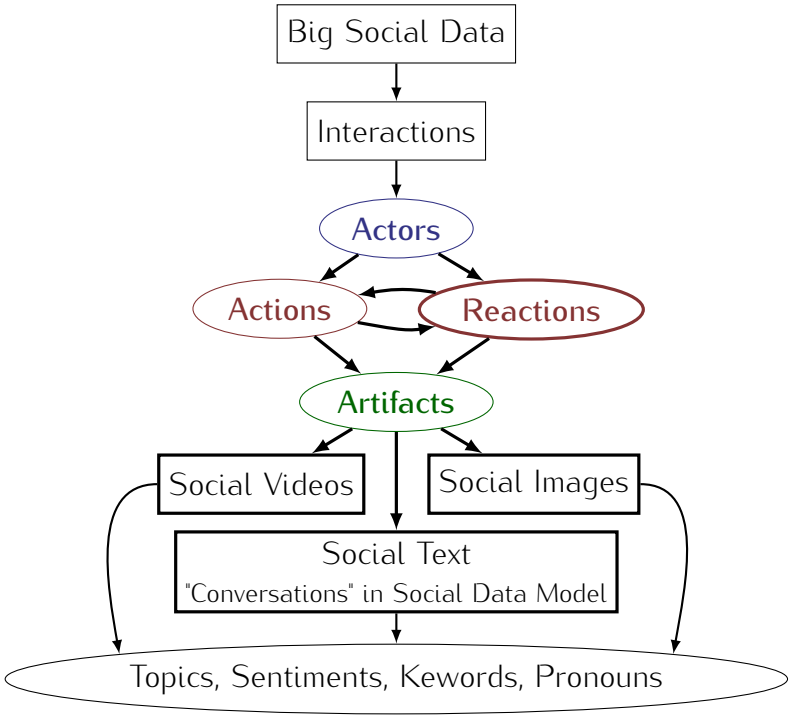


Figure 2.5: Social Interaction Model

aspects of *Topics*, *Sentiments*, *Keywords*, and *Pronouns*, which were previously introduced in the Social Data Model but limited to the concept of *Conversations*.

Furthermore, in Figure 2.5 the use of boxes vs. ellipses displays the hierarchy of concepts used within the old two-tier Social Data Model. Boxes depict a higher rank in the hierarchy, whereas ellipses depict a lower rank. The top-down flow of the Social Interaction Model emphasizes that a more fine-grained hierarchy is needed, e.g. more than two levels as in the old model. Hence, former hierarchies are only illustrated for the reader to better understand and compare the transformation of the Social Data Model.

In the Social Interaction Model, *Actors* depict **any kind of user or entity** that can be interacted with in the realm of social media. Each *Actor* consists of a unique *Location in Space* and a set of *Artifacts* which depict the actor's attributes, i.e. name, date of birth, profile picture or self-descriptive bio text. Actors depict both origin and destination of social-technical interactions.

An *Interaction* consists of **one initial *Action* and zero to many *Reactions*** that respond to the initial *Action* or one of its *Reactions*. In this model, *Actions* always occur between two *Actors*, one originating *Actor* and one receiving *Actor*. As *Actions* are always directed at other *Actors*, they follow the concept of other-directed or "transactive" actions [Berkowitz & Gibbs 1979]. Consequently, *Reactions* always originate from one *Actor* and are targeted at another *Action* or *Reaction*.

Moreover, **every single *Action* or *Reaction* spawns a newly created set of *Artifacts***. *Artifacts* can be any kind of **user-generated content**, such as text posts or com-

ments, media uploads, profile pictures, et cetera. When *Actors* want to interact with an *Artifact*, they are limited to referencing the *Actions* or *Reactions* which spawned the *Artifact* in question from a newly-created *Reaction*. By definition, *Artifacts* depict only the result from an *Action* or *Reaction*. Each individual *Artifact* consists of a certain content type and a user-generated payload. The conceptual model proposed in this thesis supports three content types, them being *Social Videos*, *Social Images* and *Conversations*. Those sets can be aggregated on the specific *Interaction*, resulting in a set of all *Artifacts* created during a certain *Interaction*. Interactions within this proposed conceptual model depict interactions in line with the theory of socio-technical interaction [Vatrapu 2010].

Through computational **analysis of individual *Artifacts***, further information regarding *Topics*, *Keywords*, *Pronouns*, and *Sentiments* can be extracted. This can be achieved by means of machine learning, e.g. for sentiment analysis [Boiy & Moens 2009, Neethu & Rajasree 2013], or more specialized techniques for text analysis [Abbasi & Chen 2007, Abbasi & Chen 2008, Abbasi *et al.* 2013].

The **formal definition of the Social Interaction Model** is illustrated in Figure 2.6. Temporal and spatial dimensions are included as a core component of the proposed model, therefore streamlining data analytics tasks in the realm of Social Set Analysis. These temporal and spatial dimensions have also been utilized in the prior theoretical model of interaction in socio-technical networks by [Suthers *et al.* 2010, Suthers & Rosen 2011], but not in the two versions of the Social Data Model. Furthermore, the *Actions* and *Reactions* in the Social Interaction Model are a specialized case of the "uptake" concept introduced by [Suthers *et al.* 2010] for the realm of Big Social Data.

Within the Social Interaction Model, the context of each *Interaction* refers to its

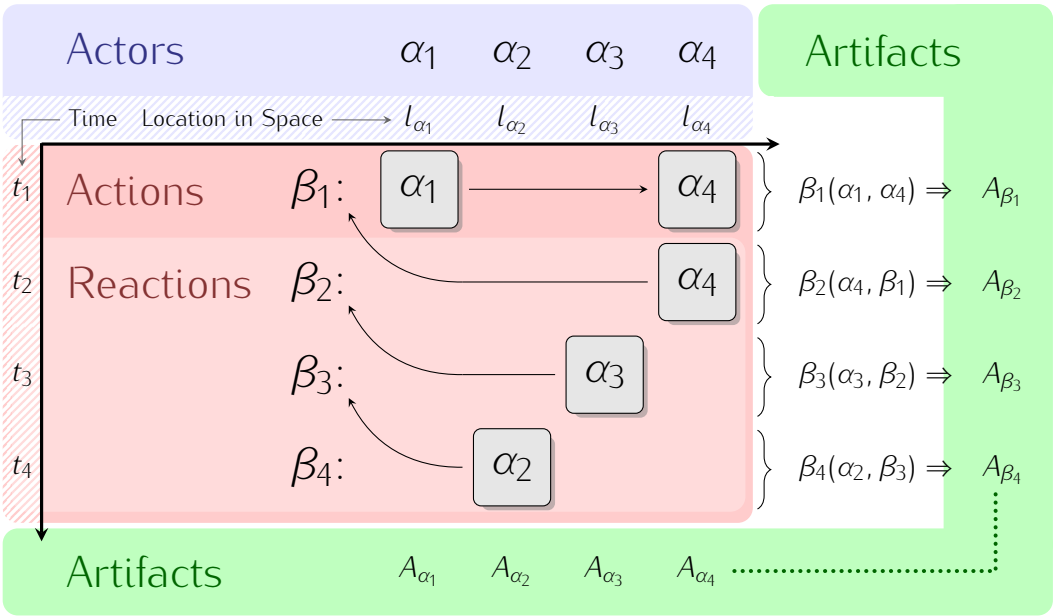


Figure 2.6: Formalization of the Social Interaction Model

Location in *Space* and *Time*, thus allowing analytics along **spatial and temporal dimensions**. *Location in Space* is provided by *Actors* who are part of the initial *Action*. As by definition every *Actor* exhibits a certain *Location in Space*. Therefore, the *Location in Space* for the whole interaction is set by the location attributes of the initial *Action* of each *Interaction*. *Location in Time* is attached to every single *Action* and *Reaction* as a conventional time stamp.

For comparison across the dimension of time, a set-based intersection of various time periods along the time axis is calculated. Similarly, for comparison across location in space, **set inclusions and exclusions** are calculated based on the dimension of space, for example a list of relevant Facebook walls. Based on this established procedure, we can design a theoretical model that fits to analytical Social Set Analysis methodologies as they are employed in various case studies.

The reasons for a conceptualization of the Social Interaction Model as a **generalized theoretical model of interactional Big Social Data** are manifold. First, it is to extend upon the core ideas put forward through the Social Data Model. Secondly, it is to apply learnings from a series of set-based Visual Analytics studies performed during my PhD project using the Social Set Visualizer software tool. And lastly, it is to improve the existing theoretical data model in which the Social Set Analysis approach to Big Social Data Analytics is grounded.

Accordingly, the Social Interaction Model introduces a set-based definition of *Interactions* and the resulting *Artifacts* within Big Social Data. It proposes a **two-dimensional theoretical framework** based on the two concepts of location in space and location in time. Such a framework provides additional theoretical coherence with the empirical application of set-based analytics as presented throughout recent publications by our research group [Vatrapu *et al.* 2014] and the multiple publications on this topic included in this dissertation. Therefore, the concepts of the previously utilized Social Data Model are refined, formalized and simplified.

Learnings from various computational implementations during the course of this thesis are incorporated in the Social Interaction Model. **Key differences** between the existing Social Data Model and the Social Interaction Model present the inclusion of non-textual artifact content types, the unification of a previously bipartite Social Data Model, the depreciation of an empirically vague notion of *Activities*, and the improved interoperability between different sources of Big Social Data.

In order to simplify the theoretical model for practical utilization in Big Social Data Analytics, *Actions* and *Reactions* are **limited to one-on-one relationships**. This is grounded in the nature of the most widely used social media platforms, which also limit the structured flow of discussions to one Action or Reaction at any single point in time, therefore not natively supporting one-to-many *Actions* or *Reactions*.

Due to this restriction, a **tree-like data structure** emerges from the formalization of the Social Interaction Model as one initial *Action* exists for each *Interaction*, with one-on-one *Reactions* developing a tree structure from this initial *Action*. This structure of the model reflects the structure of the Big Social Data that was analyzed in the course of the different projects und publications that were part of this thesis. Due to the filtering along the spatial and temporal dimensions of the data at hand,

initial reference points for the analyzed interactions are established on which the tree structure crystallizes.

2.4 Data Collection

The collection of Big Social Data from Facebook has been refined and streamlined in recent years. Previous work by members of our research group, most notably Abid Hussain, pioneered this approach through the creation of software tools such as the *Social Graph Analytics Tool (SOGATO)*, released in 2011, and the *Social Data Analytics Tool (SODATO)*, released in 2014.

While the initial version of the Social Set Visualizer relied on datasets fetched through SOGATO and SODATO, the third and final release of the Social Set Visualizer software tool contained built-in data crawler which directly integrates this functionality. Thereby, the time-consuming importing and exporting of file-based datasets is streamlined. This radically simplifies usage of the Social Set Visualizer software tool for research purposes.

2.4.1 Social Graph Analytics Tool (2011)

The Social Graph Analytics Tool (SOGATO) was presented in 2011, depicting the first publication of a Big Social Data IT artifact with special focus on social media data from public pages in the Facebook ecosystem [Hussain & Vatrapu 2011]. It supported data fetching of multiple Facebook pages and showed extensive statistics on the fetched data, as illustrated in Figure 2.7. The researchers showcased its utility by

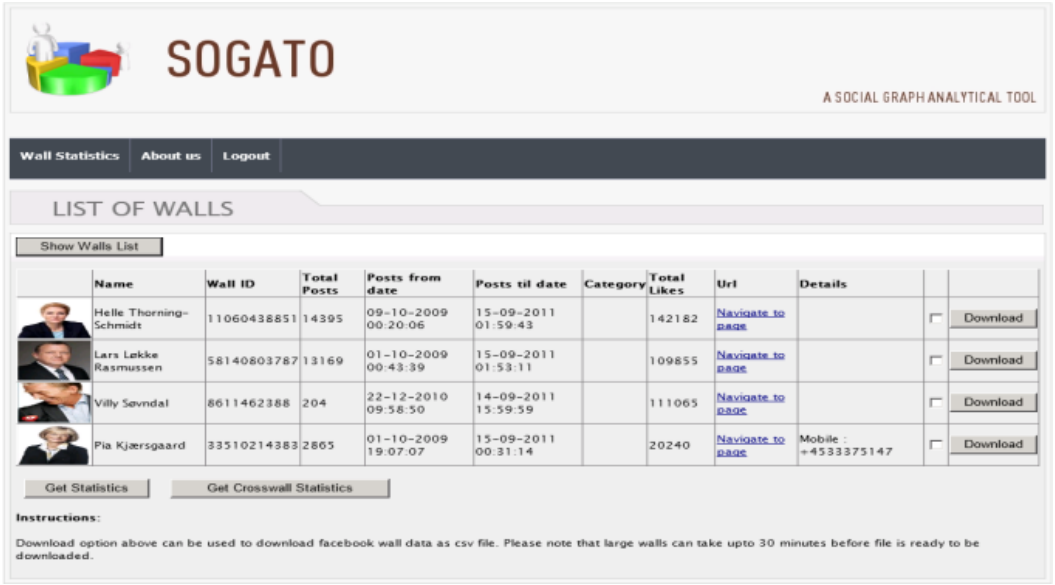


Figure 2.7: Social Graph Analytics Tool [Hussain & Vatrapu 2011]

means of a demonstrative case study on the 2011 parliamentary elections in Denmark. Several descriptive statistics such as counts of posts, likes and simple cross-wall comparison between pages are provided in the software tool. Furthermore, users of the Social Graph Analytics Tool could download the Big Social Data for additional analysis through advanced methods.

2.4.2 Social Data Analytics Tool (2014)

The Social Data Analytics Tool (SODATO) was first presented in 2014 to an international audience [Hussain & Vatrapu 2014a, Hussain & Vatrapu 2014b, Hussain *et al.* 2014]. The Social Data Analytics Tool represents an iteration of the Social Graph Analytics Tool. It describes the building step towards the vision of a fully integrated data fetching and analytics platform for Big Social Data from the Facebook platform.

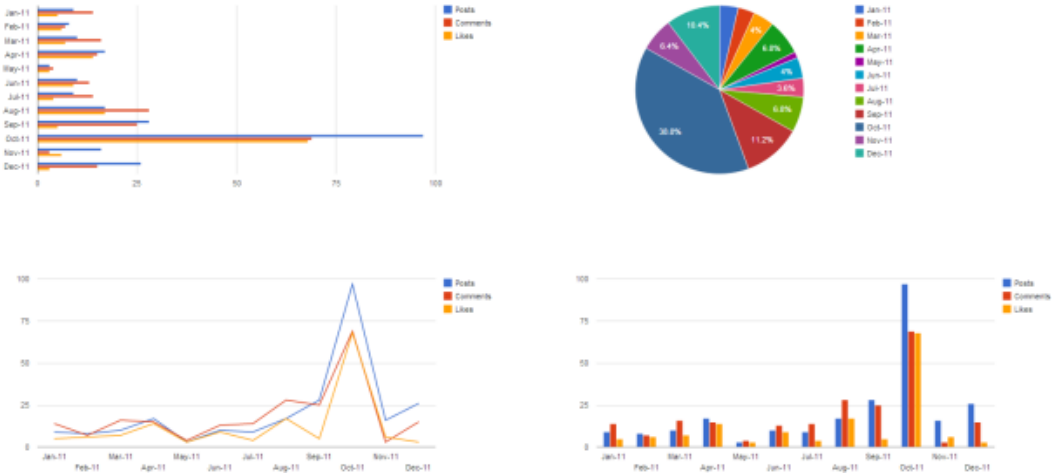


Figure 2.8: Social Data Analytics Tool [Hussain & Vatrapu 2014a]

The Social Data Analytics Tool provides a **data fetching interface** similar to Social Graph Analytics Tool and further introduces a dashboard-style visualization of relevant metrics regarding the fetched Big Social Data. It was implemented on a Microsoft platform using the Microsoft SQL Server as a storage backend. Figure 2.8 showcases the Social Data Analytics Tool. However, it does not provide means for performing interactive Visual Analytics of the downloaded Big Social Data.

2.4.3 Built-in Data Crawler in SoSeVi 3 (2017)

In 2017, a **built-in Facebook data crawler** was added to the third version of the Social Set Visualizer, which is detailed in **Publication IV** [**Flesch** *et al.* 2017] of this dissertation. The integration of a data crawler into the Social Set Visualizer provides the overarching benefit that data conversion and cleanup tasks are massively simplified, and removed as source of potential errors. Data is fetched through the Facebook

graph API and stored in a custom database schema according to the theoretical data model described earlier.

Before this integration was made, the data for versions one and two of the Social Set Visualizer software tool was fetched using the Social Data Analytics Tool [Hussain & Vatrapu 2014b]. Data collected through the Social Data Analytics Tool is saved as independent files and needs to be imported into the Social Set Visualizer before it can be displayed. In order to **remove this complexity** from the Big Data Value Chain, a built-in Facebook data crawler was designed and implemented directly within the Social Set Visualizer.

2.4.4 List of Datasets

Lastly, a list of datasets is provided including all Facebook datasets that have been collected and analyzed over the course of this PhD project. Table 2.1 depicts six individual Facebook datasets and the peer-reviewed research publications which arose from these datasets. The Facebook datasets from before 2017 have been downloaded using the Social Data Analytics Tool. The datasets from 2017 and 2018 have been fetched using the built-in Facebook crawler of the Social Set Visualizer software tool.

Publication	Total Facebook Data Points
[Vatrapu *et al.* 2015]	180,000,000
[**Flesch** *et al.* 2018]	138,989,847
[**Flesch** *et al.* 2015b]	89,654,347
[**Flesch** *et al.* 2017]	15,000,000
[Boldt *et al.* 2016]	10,000,000
[Hennig *et al.* 2016]	7,532,000

Table 2.1: Overview of Facebook datasets collected in this PhD project

2.5 Analytical Processes

The analytical processes of the work presented in this dissertation revolve around the interactive definition of sets based on **criteria of time, location in space, and other attributes such as Actors, Actions or Artifacts** that stem from the Social Interaction Model. Based on these set definitions, the Social Set Visualizer software tool is utilized to compute and visualize all relevant set intersections from the large-scale datasets which are stored in its database. Thereby, the researcher is able to observe and identify interesting phenomena in the Big Social Data at hand. Over the course of multiple iterations, the set definitions can be refined in order to capture the essence of the researched phenomenon.

Advanced analytical processes such as identifying important actors, quantifying who associates with whom, who produces what kind of content, and identifying cohesive clusters of actors can also be performed using this set-based approach.

In order to **identify important actors in social media**, we can use the Social Interaction Model to quantify the number of reactions of to actions that are initiated by each actor. By filtering the types of actions and reactions, the analysis can be focused on a specific facet of the actor's behavior. An example of this would be our still unpublished study on the topic of fake news in the 2016 US elections, in which text-only artifacts containing hyperlinks were identified, then the hyperlinks were extracted, and finally they were compared to a known-bad list of fake news websites. Thereby, the automatic posting of fake news links by social media bots can be quantified.

In order to quantify, not just visualize, **who associates with whom**, the analytical process is as follows. First, set-based definitions are created based on the metrics that should be explored. For example if we want to assess the political leaning of sports teams, we fetch data for a sports team and data for all political parties. Then, a pairwise set intersections between sports teams and political parties are calculated. These set intersections are quantified by the software tool, and allow researchers to identify what percentage of each sports team is associating with each political party.

In order to **identify who produces what kinds of content**, a simple approach is sufficient. It could be based on filtering and aggregating different types of artifact content that is created by authors. Depending on the research question, an interactive software tool is not really needed due to the limited scope of the research problem.

In order to **identify cohesive clusters of actors**, for example in a study on elections and social media, actors can be grouped in month-based sets in order to quantify party loyalists. Once the set (or cluster) of loyalists is identified, we can both compare artifacts produced of loyalists vs. non-loyalists, and also what kind of artifacts are produced by loyalists on the other parties' pages. Furthermore, due to the set-based units of analysis, it is straightforward to perform studies akin to churn analysis in businesses, in which the movements of groups of actors within social media are quantified. Thereby, the effects of marketing and election campaigns can be measured and analyzed.

2.6 Summary

This chapter introduced the research methodology of my PhD project. First, it described Action Design Research and its IT-dominant version of the building, intervention and evaluation loop which was chosen for this dissertation. Second, the two analytical techniques of Social Set Analysis and Event Study Methodology were presented. They depict the foundational components of the set-based approach to Big Social Data Analytics utilized in this thesis. Third, the existing theoretical model of Big Social Data, the Social Data Model, was introduced. Moreover, its advancement, the Social Interaction Model, was described and reasons for its creation were outlined. Fourth, the data collection pipeline of this PhD project was described and the Facebook data crawler which is incorporated into the Social Set Visualizer software tool was presented. Lastly, a list of six large-scale Facebook datasets utilized in this

PhD project was given, providing the baseline for the evaluative case studies, and the different analytical processes were highlighted.

Chapter 3

Design

This chapter details the design of the Social Set Visualizer software tool. For this, first, the target audience of the Social Set Visualizer is defined, in line with the stakeholders of the conducted case studies. Second, the design goals and objectives of the Visual Anayltics tool are outlined, providing the means of reaching a high level of usability for the software tool. Third, the three major user interfaces of the Social Set Visualizer are introduced, namely a browser-based dashboard, a visual query builder, and a textual query language. Furthermore, the state of the art in set-theoretical visualizations is outlined. This chapter is thus concluded by the presentation of area-proportional visualizations of large-scale sets and set intersections, that have been designed over three versions of the Social Set Visualizer.

3.1 Target Audience

The **target audience** for the Social Set Visualizer software tool are end-users both from academia and industry. As showcased in the introductory chapter, researchers from the field of data science utilize the Social Set Visualizer in order to generate **novel insights** from Big Social Data that depict a publishable contribution to the state of the art. Furthermore, industry practitioners such as social media analysts utilize the software tool for the generation of **actionable insights** that assist their decision making process.

These two target audiences are in accordance with the various stakeholders of the descriptive and predictive case studies which are presented in section 5.1 and section 5.2 of this dissertation.

3.2 Design Goals & Objectives

In line with the first research question, the overall design goal for the Social Set Visualizer is to devise an interactive Visual Analytics software tool that incorporates the Social Set Analysis approach to Big Data Analytics and effectively visualizes sets and set relations. Thereby, the work presented in this PhD project should **enable the answering of research questions and the discovery of novel patterns** in Big Social Data. Furthermore, the second research question outlines the goal of identifying software design requirements for the visualization of large-scale sets.

Based on these two underlying design goals derived from the research questions and also on the presented target audience from industry and academia, a sensible

set of **design objectives** needs to be selected in order to guide the subsequent development of the user interface for the Visual Analytics software tool.

In this context, design objectives depict "the functional and non-functional qualities of a design, [..] that guide the design process and measure results" [Spacey 2018]. In the field of human-computer interaction, design objectives are utilized in order to **guide the design of high quality user interfaces** [Issa & Isaias 2015].

Therefore, **usability depicts the core design objective** for the Social Set Visualizer software tool which is designed and developed over the course of this PhD project. Usability consists of five design objectives, namely efficiency, learnability, memorability, user satisfaction, and error handling [Scholtz 2004]. These five design objectives guide the development of the user interface for the Social Set Visualizer.

Efficiency depicts an essential design objective, as the user interface needs to immediately convey important information and offer an optional deep dive using additional details on demand. The user interface of the Social Set Visualizer displays Big Social Data, predominantly from Facebook. Therefore, it consists of a combination of multiple visualizations. Each visualization needs to efficiently highlight specific features of the underlying data. This allows the dashboard as a whole to be kept clean and organized, preventing it from becoming too complex for the user. The **detail on demand principle** improves efficiency insofar that it first presents an easily understandable overview to the user, which can then be enhanced with additional details if the user chooses to do so. Therefore, the initial visualization can be quickly processed on a visual and intellectual level. Subsequently, once the user explicitly demands it, the level of detail shown in the visual analytics tool is increased. Using this approach, large-scale datasets can be easily processed by the user. Hence, it leads to greater productivity when using a Visual Analytics software tool.

Furthermore, the design objective of **learnability** can be formalized in the learning curve of the software tool. *Are users able to complete the tasks they set out to achieve and how many errors happen along the way? Is the tool comprehensible, quick to understand, and intuitive?* For end users, learnability depicts an important non-functional requirement. The Visual Analytics software tool should be thus designed in a way that enables users to work with the dashboard without any prior briefing or formal training on how to use it, while minimizing the frequency of errors.

Additionally, the user interface needs to be memorable for the user, as research has shown that improvements in **memorability** of a user interface also positively influences its learnability [Van Welie & Trætteberg 2000, Hartmann *et al.* 2008]. This can be achieved by following design conventions and guidelines which are already familiar to the target audience, thereby helping to reduce the cognitive load on the user. Hence, memorability depicts an influential design objective in order to reach good usability for the Social Set Visualizer software tool.

Moreover, **user satisfaction** depicts another important design objective in order to reach a high level of usability. Once basic user needs such as functions and features are met by the IT artifact, user satisfaction can be influenced by a persistently positive user experience within the application. A great user experience can be created through user-centric design. The approach of user-centric design emphasizes that

"the role of the designer is to facilitate the task for the user and to make sure that the user is able to make use of the product as intended and with a minimum effort to learn how to use it" [Abras *et al.* 2004]. Therefore, when designing the user interface, a focus is put on optimization of the user experience. Furthermore, the design of the software tool should follow the principle of least astonishment, which was first formulated by [James 1987], and later applied to modern usability testing [Isaksen & Bertacco 2006]. It states that good user interfaces should never surprise the user, and that the user should know how to use the tool based on intuition. The principle of least astonishment goes also hand in hand with the previously outlined design objective of learnability, which further outlines the interdependence of the presented design objectives.

Lastly, the essential design objective of **error handling** in user interfaces concludes the concept of usability. In order to both prevent and capture human errors, but also communicate application errors, a suitable approach to error handling needs to be utilized by the IT artifact. Part of this approach is the concept of graceful degradation, where the occurence of errors is limited to certain parts of the user interface, without negatively affecting or even shutting down the application as a whole.

Together, the five presented design objectives of efficiency, learnability, memorability, user satisfaction, and error handling depict the means of reaching a high level of usability for the developed Social Set Visualizer software tool.

3.3 User Interfaces of the Social Set Visualizer

In this section, the central user interfaces of the Social Set Visualizer will be introduced and detailed from a design and usability perspective. These user interfaces have been designed based on the previously presented theoretical foundations of usability. The presented iterations allow the users to efficiently conduct analyses based on Social Set Analysis methodology, overcoming different data formats and structures as well as the challenge of visualization large-scale sets and set intersections.

3.3.1 Browser-based Visual Analytics Dashboard (2016)

The central user interface of the Social Set Visualizer is a browser-based Visual Analytics dashboard, through which the user can interact with the datasets. The design of the user interface has evolved over several versions of the Social Set Visualizer.

The first version of the user interface, presented in **Publication II** [**Flesch** *et al.* 2015a] of this dissertation, is illustrated in Figure 3.1. The dashboard interface depicts several areas, most prominently a large line chart with overall activity visualization and, directly below, a timeframe selection brush. At the bottom of the interface, a word cloud for the selected time frame is visualized to interpret content. On the right hand side, the set-based visualization according to the Social Set Analysis methodology is shown. In the bottom right corner, results are shown from a language detection heuristic that is applied to the conversion content of the selected timeframe.

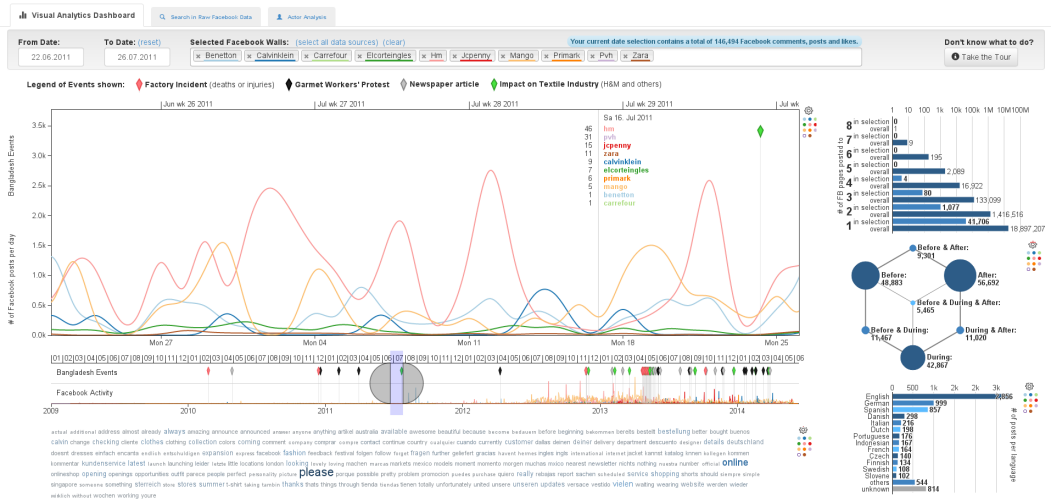


Figure 3.1: Browser-based Visual Analytics Dashboard in SoSeVi 1 (**Publication II** [**Flesch** *et al.* 2015a])

In this first version of the dashboard, the user is able to interact with the Social Set Visualizer through various means. The Social Set Visualizer allows the user to select different Facebook walls by using a text input field at the very top of the user interface. It supports both textual input and drop-down selection of Facebook walls for analysis purposes. Furthermore, the time period of interest can be selected either by text input in the top right input field, by selection using the calendar widget which is attached to the input field, or through use of the visual time slider underneath the main visualization.

Figure 3.2: Browser-based Visual Analytics Dashboard in SoSeVi 2 (**Publication III** [**Flesch** *et al.* 2016])

The second version of the user interface, presented in **Publication III** [Flesch *et al.* 2016] of this dissertation, is illustrated in Figure 3.2. In this version, the time selection interface has been placed at the top of the dashboard. Furthermore, the set-theoretical visualization was moved to the left hand side, while the main activity visualization and the word cloud were reduced in size and placed at the right.

3.3.2 Visual Query Builder (2017)

In the third version of the Social Set Visualizer, the user interface was significantly streamlined and simplified in order to allow a more efficient implementation of research studies using the **novel, set-based approach to Big Social Data Analytics**. This is detailed in **Publication IV** [Flesch *et al.* 2017] of this dissertation.

Due to incorporation of the Facebook data crawler, the third version of the Social Set Visualizer interface was transformed into a **wizard-like dialog** consisting of three steps, which is illustrated in Figure 3.3.

The first step of this dialog is to **select the Facebook walls** of interest. In this step, both a checkbox-style selection of relevant Facebook walls and a textual input field for type-in search are offered to the user. After the user has selected relevant Facebook walls for the set-based analysis, a dynamic **selection of the time periods** appears in the second step. For each of the selected Facebook walls, a sparkline-style timeline of the available, previously fetched data is displayed. The user can then use a time selection tool for each of the selected Facebook walls in order to choose either the full Facebook wall for analysis or only a certain time period. This step depicts the defition of sets along the dimensions of space and time. Furthermore, two buttons offer additional functionality. The *Duplicate* button copies the currently

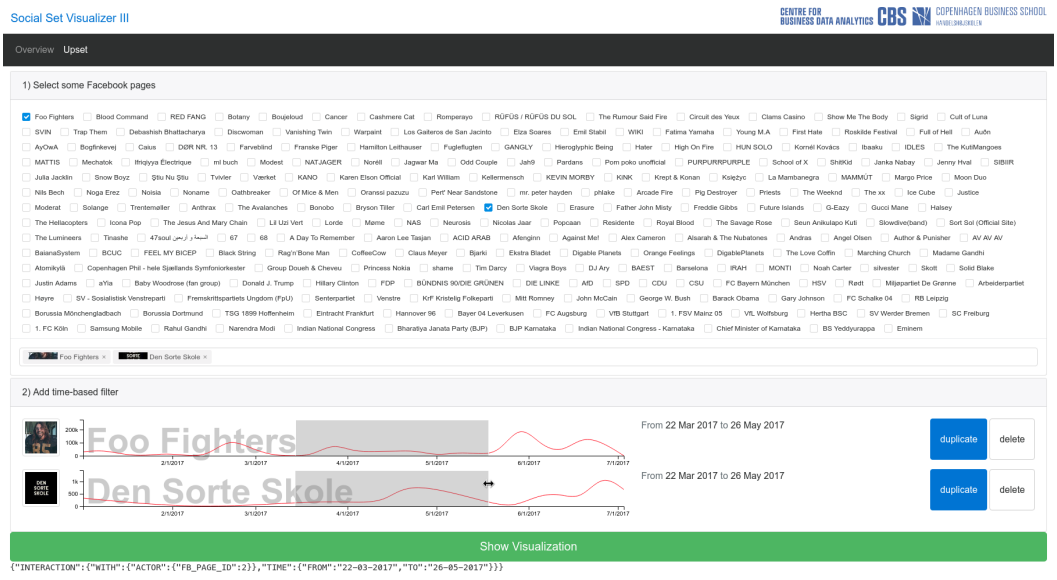


Figure 3.3: Visual Query Builder in SoSeVi 3 (**Publication IV** [Flesch *et al.* 2017])

selected set and appends it to the end of the selection. The *Delete* button removes the selected set. Once the selection of time- and space-based sets from the Big Social Data is concluded, the user can press the *Show Visualization* button, after which the **set visualization is generated** and displayed to the user.

3.3.3 Textual Query Language (2018)

In 2018, the third version of the Social Set Visualizer was extended by a **textual query language**. This textual query language provides a textual user interface which enables the user to perform structured studies using the Social Set Visualizer software tool, without having to rely on the visual user interface for the selection and definitions of sets.

The textual user interface is illustrated in Figure 3.4. It utilizes a specific query language, the **Social Set Query Language**. The Social Set Query Language is a domain-specific textual query language that implements the theoretical model of Big Social Data, the Social Interaction Model, which was presented in section 2.3.4 of this thesis. It provides an additional user interface for the Social Set Visualizer software tool to perform set-based analytics through querying and retrieving data from a corpus of Big Social Data. As a textual query language, it depicts the API for interacting with the database backend of the Social Set Visualizer. The Social Set Query Language was first presented as part of the third version of the Social Set Visualizer in **Publication IV** [**Flesch** *et al.* 2017] of this dissertation.

Design and development of the Social Set Query Language is motivated by the overarching objective of this dissertation to provide means of **simplifying Social Set Analysis tasks for researchers and practitioners** alike. During iterations on this PhD project and the IT Artifacts involved, it became apparent that an expressive query language is needed to effectively formulate set queries. This query language should be abstracted from the actual data storage backend and be based on a suitable theoretical model of Big Social Data. In Computational Social Science, SQL is utilized by researchers for analytical workflows of Big Social Data. When using a relational database, it is universally queried using SQL, a textual query language. Therefore,

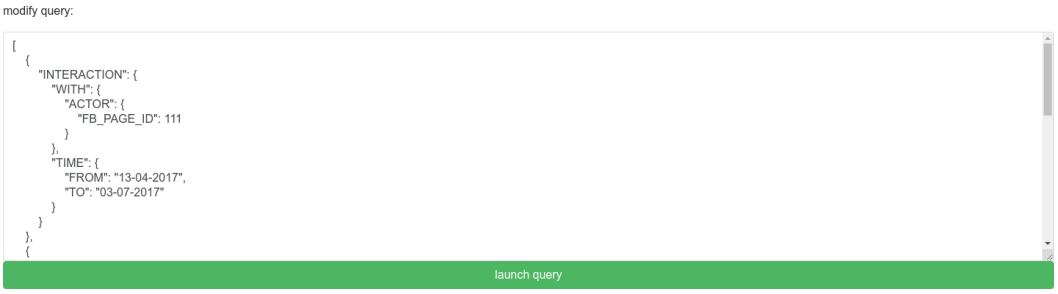


Figure 3.4: Textual query interface based on the Social Set Query Language in SoSeVi 3 (**Publication IV** [**Flesch** *et al.* 2017])

utilization of a normalized database designed on a theoretical model of social data is the sensible choice for such an undertaking.

The highly structured and condensed data stored according to the Social Interaction Model provides normalized data with a relatively small storage footprint. Queries against the database need to utilize expressive SQL and thus, can become very verbose. This in turn creates a **significant amount of friction**, which can be reduced through introduction of a domain-specific textual query language.

For the purpose of Social Set Analysis, row-based data needs to be first converted into a set-based representation in order to perform set operations on the previously queried sets. Thus, a **simplification of syntax** is required to streamline the querying of data within the Social Set Analysis methodology.

Recent studies underline that **domain-specific query languages** have been created for various research fields in order to simplify access to data and to streamline the conversion of user input into specific outputs [Van Deursen *et al.* 2000, Atkins *et al.* 1999]. The transformation of SQL into domain-specific query languages or even into query languages for other non-relational databases such as Apache Spark and NoSQL databases has been demonstrated in various research projects [Armbrust *et al.* 2015, Liu *et al.* 2016]. Furthermore, other examples for the successful application of domain-specific query languages to research problems in data mining and Big Data Analytics exist [Goebel & Gruenwald 1999].

Domain-specific query languages have been proposed for the field of Social Network Analysis in order to simplify analytical tasks based on the network structure which is stored as a graph in a database [Seo *et al.* 2013]. An equivalent domain-specific query language for purposes of Social Set Analysis does not exist at this time, therefore its creation as part of this PhD project is **important for further dissemination of the set-based approach**.

Therefore, the Social Set Query Language presented in this thesis depicts the **first domain-specific query language for the field of Social Set Analysis**. The Social Set Query Language converts a simple JSON-style set definition into SQL that queries data from the database. Therefore, the formulation of queries within the Social Set Analysis framework is decoupled and abstracted from the requirement of expert knowledge in SQL.

Listing 3.1 provides an example query of the Social Set Query Language which fetches all actors that interact with a certain Facebook page over three distinct intervals. The first set refers to a time interval from 1st of April 2017 to 30th of April 2017, the second set to a time interval from 1st of May 2017 to 31st of May 2017, and the third set to a time interval from 1st of June 2017 to 30th of June 2017. Subsequently, the query engine fetches all three sets and performs a set intersection of the interval data. Finally, the cardinality of the set intersection is calculated and returned.

Listing 3.1: Example of a set-based query in the Social Set Query Language

```
[
   {
      "INTERACTION":{
         "WITH":{
            "ACTOR":{
               "FB_PAGE_ID":201
            }
         },
         "TIME":{
            "FROM":"01-04-2017",
            "TO":"30-04-2017"
         }
      }
   },
   {
      "INTERACTION":{
         "WITH":{
            "ACTOR":{
               "FB_PAGE_ID":201
            }
         },
         "TIME":{
            "FROM":"01-05-2017",
            "TO":"31-05-2017"
         }
      }
   },
   {
      "INTERACTION":{
         "WITH":{
            "ACTOR":{
               "FB_PAGE_ID":201
            }
         },
         "TIME":{
            "FROM":"01-06-2017",
            "TO":"30-06-2017"
         }
      }
   }
]
```

The formal definition of textual queries in the Social Set Query Language can be expressed in the JSON schema format. Listing 3.2 showcases this definition.

Listing 3.2: Formal JSON schema definition for Social Set Query Language

```
{
  "$schema":"http://json-schema.org/draft-07/schema#",
  "type":"array",
  "items":{
    "title":"The Social Sets are defined in this Array",
    "properties":{
      "INTERACTION":{
        "title":"The Interaction Schema",
        "properties":{
          "WITH":{
            "title":"The definition of a Location in Space",
            "properties":{
              "ACTOR":{
                "title":"The Actor from Social Interaction Model",
                "properties":{
                  "FB_PAGE_ID":{
                    "type":"integer",
                    "title":"The Facebook Page ID"
                  }
                }
              }
            },
            "TIME":{
              "title":"The definition of a Location in Time",
              "properties":{
                "FROM":{
                  "type":"date",
                  "title":"Start date of the time interval",
                  "pattern":"^([0-9]{2}-[0-9]{2}-[0-9]{4})$"
                },
                "TO":{
                  "type":"date",
                  "title":"End date of the time interval",
                  "pattern":"^([0-9]{2}-[0-9]{2}-[0-9]{4})$"
                }
              }
            }
          }
        }
      }
    }
  }
}
```

The Social Set Query Language is executed through an API which is situated on top of the server-side database that stores the actual data. This way, the user-controlled, client-side component of the Social Set Visualizer software tool is able to send user-formulated queries to the server-side database which are in line with the Social Set Analysis methodology. Therefore, the client requires no deeper knowledge about the underlying data structures of the server-side database or the intricacies of optimizing the SQL queries that are involved. Thus, the client only needs to pose queries using the Social Set Query Language, and all resulting **complexity is handled by the server-side** part of the Social Set Visualizer.

The query language is **implemented through database functions and stored procedures**, which have been streamlined and optimized for the database setup at hand. All queries are performed on top of a database schema that implements the Social Interaction Model. Based on this, a reference implementation of the Social Set Query Language for Big Social Data from Facebook as used in the Social Set Visualizer has been created.

3.4 Visualization of Sets

The visualization of sets poses **two significant challenges** to the design of the Social Set Visualizer software tool that is presented in this thesis.

First, the **visual comparability between different sets** is an essential requirement for the Social Set Visualizer, as otherwise it would fail its very purpose as an interactive Visual Analytics tool. Therefore, some sort of area-proportional visualization is required, in order to achieve visual comparability.

Second, the **visualization of a large number of sets** poses a serious problem, as the set-based approach to Big Social Data Analytics usually results in many different sets that need to be displayed. In this context, a reduction of the number of total sets, and thereby the complexity of the set visualization, is oftentimes not possible without potentially losing valuable analytical insights.

This section will provide an overview about the state of the art in set visualization and how these two significant challenges are overcome during the design of the Social Set Visualizer.

3.4.1 Euler and Venn Diagrams

Both Euler and Venn diagrams depict the most prominent **set-theoretical visualizations**. They are frequently utilized in order to visualize two or more different sets and to visually highlight their set intersections through the use of multiple closed shapes.

Euler diagrams are widely attributed to Leonard Euler (1707–1783), who first presented a set-based visualization using intersecting circles to the academic audience [MacQueen 1967]. Euler diagrams visualize all naturally occuring set intersections within the provided sets. Figure 3.5 displays an Euler diagram which illustrates seven different sets and the intersections between some of those sets.

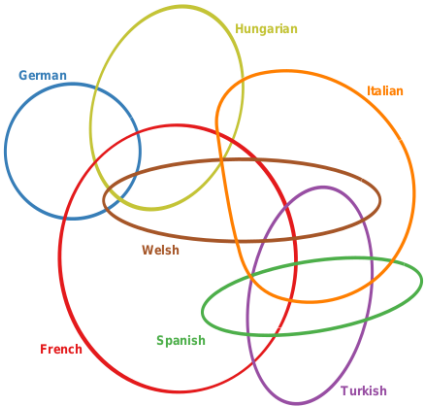


Figure 3.5: Euler Diagram [Rodgers *et al.* 2015]

In 1880, John Venn (1834-1923) introduced the **Venn diagram**, a special case of the Euler diagram, that displays all possible set-theoretical intersections, including empty intersections which might not naturally occur between different sets [Venn 1880, Ruskey & Weston 1997], as illustrated in Figure 3.6.

The two diagram types utilize an area-based visualization of the cardinality of sets. However, an **area-proportional approach** to set visualization is needed in order for the viewer to visually discern between sets of different sizes. Previous work has emphasized that the drawing of area-proportional Euler and Venn diagrams poses significant difficulties [Chow & Ruskey 2003, Chen & Boutros 2011].

Even though Euler and Venn diagrams can be area-proportional under some circumstances, they exhibit **substantial practical limitations** when visualizing as few as three different intersecting sets.

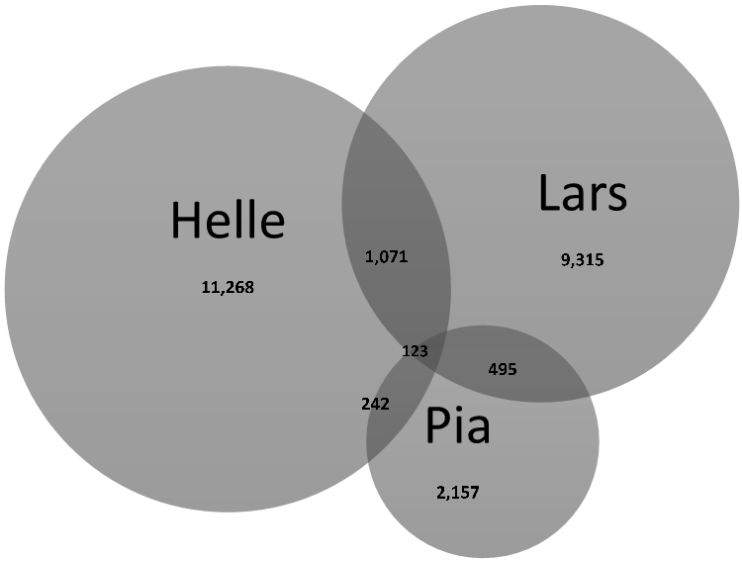


Figure 3.6: Three-set non-proportional Venn Diagram [Vatrapu *et al.* 2015]

Figure 3.6 showcases the problem of creating a static representation of an area-proportional Venn diagram using circles, taken from an earlier publication of our research group [Vatrapu *et al.* 2015]. During close examination of the displayed intersection areas, it becomes apparent that the area of $|(Helle \cap Pia) \setminus Lars| = 242$ is significantly larger than twice the area of the three-way intersection at the center of the Venn diagram, which is expressed as $|Helle \cap Pia \cap Lars| = 123$.

Moreover, the **visualization of a large number of sets** using an Euler or Venn diagram remains challenging. A Venn diagram with six different sets is visualized in Figure 3.7, an illustration that depicts the intersections between genomes of the banana and five other species [D'hont *et al.* 2012]. While it highlights the creative possiblities with regard to the use of shapes during creation of a Venn diagram, both utility and contained information are severely limited from a Visual Analytics perspective. Due to the lack of area-proportionality, a visual comparison of the cardinality of different sets is nearly impossible. These practical difficulties in visualizing area-proportional Venn diagrams became apparent during development of the first version of the Social Set Visualizer. As a Visual Analytics software tool, the area-proportionality of set-theoretical visualizations is an **essential requirement** for the utility and usability of the application.

Therefore, the classical **Venn and Euler diagrams** are neither a good fit for this use case nor for set-based visualization in the Social Set Visualizer where an even larger number of sets needs to be displayed.

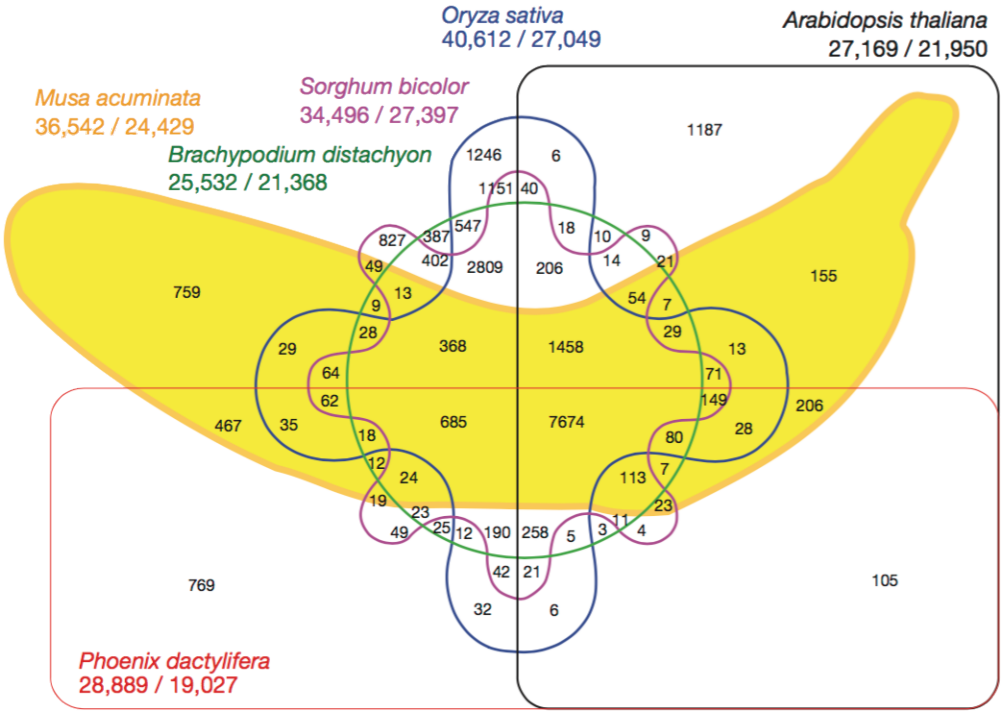


Figure 3.7: Six-way Venn Diagram of Banana Genome [D'hont *et al.* 2012]

3.4.2 EulerAPE (2012)

In 2012, an evolution in the area-proportional visualization of Euler and Venn diagrams was presented with the EulerAPE project [Micallef & Rodgers 2012]. EulerAPE resolves the complicated problem of **area-proportional visualization of three-set Venn diagrams** through the use of ellipses instead of circles. Figure 3.8 showcases two examples of these ellipsis-based area-proportional set visualizations.

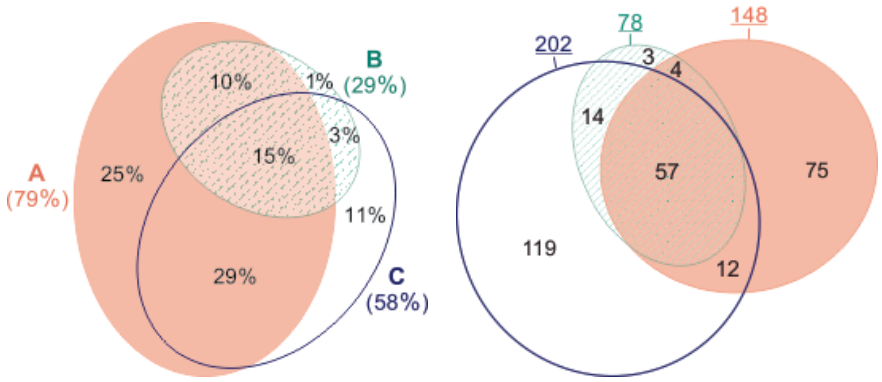


Figure 3.8: Area-proportional EulerAPE Diagrams [Micallef & Rodgers 2012]

Even though EulerAPE is an important contribution to three-way area-proportional Venn diagrams, the visualization of area-proportional Venn diagrams that consist of more than three sets **remains a significant problem**.

3.4.3 "Exploded" Venn Diagrams in SoSeVi 1 (2015)

In 2015, during design of the first version of the Social Set Visualizer, the **limitations of Euler and Venn diagrams** for interactive Visual Analytics of Big Social Data became increasingly apparent.

In this early version of the Social Set Visualizer, the set-theoretical visualization was concerned with displaying sets consisting of social media *Actors* from the *Before*, *During*, and *After* periods in the analyzed Big Social Data. This initial **three-set visualization approach** is based on the one-dimensional application of a set-based Event Study Methodology which was outlined in section 2.2.2.

However, as the three-set visualization is embedded in an interaction Visual Analytics dashboard, a **high level of usability is required**. This includes the visual comparability of sets with sizes that potentially differ by several orders of magnitude, and furthermore, the visual consistency of the displayed labels and shapes of the set intersections.

Prompted by these requirements, **an innovative approach to area-proportional Venn diagrams** was devised during the creation of the first version of the Social Set Visualizer, which is presented in **Publication II** [**Flesch** *et al.* 2015a] of this dissertation. Therefore, this set visualization formally depicts an Euler diagram variant

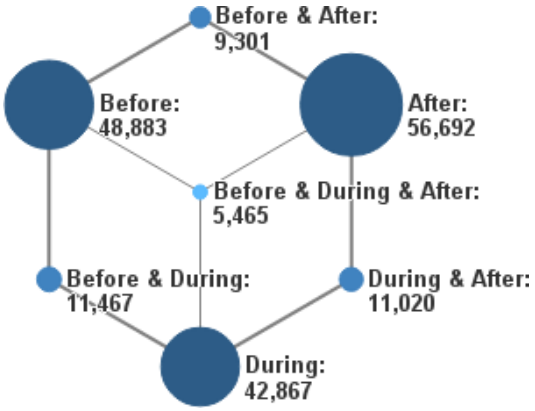


Figure 3.9: "Exploded" Venn Diagram in first version of Social Set Visualizer

that consists of a crossover between affiliation networks and circle-based Venn diagrams [Alsallakh *et al.* 2016]. The set intersections within the Venn diagram were moved into a rhombus-shaped grid with connections displayed as graph-style edges, and area-proportional circles were utilized for displaying the exact, non-distorted area of each set and set intersection. In the "exploded" Venn diagram, value labels always remain at the same position, which increases memorability of the interface and speeds up the reading of values from the set visualization. Furthermore, the center points of each circle are also fixed and remain in the same spot, thereby increasing usability when values and set sizes change. Due to the interactive nature of the software tool, the visual consistency makes it easier for the user to take a quick glance at the visualization in order to read the information that is needed.

This modification to the traditional Venn diagram provided **significant utility for users of the Social Set Visualizer** software tool, as users appreciated the visual comparability of set cardinalities, and visual consistency of both labels and shapes in the visualization. Figure 3.9 showcases this **"exploded" Venn diagram**. It signifies the challenge of visualizing sets and set intersections with largely varying cardinalities, as it is often the case when using the set-based approach to Big Social Data Analytics.

3.4.4 Linear Diagrams (2014)

In 2014, the visualization of sets using linear diagrams was generalized by [Chapman *et al.* 2014]. Linear diagrams depict set-based visualizations which are created "using **straight line segments**, with line overlaps corresponding to set intersections" [Rodgers *et al.* 2015].

Comparative studies conclude that set-theoretic tasks can be performed **significantly better using linear diagrams** instead of Euler and Venn diagrams [Gottfried 2015, Rodgers *et al.* 2015].

Figure 3.10 shows a linear diagram displaying seven different sets and their intersections. From the diagram, it is easy to recognize which sets are intersecting,

Figure 3.10: Linear Diagram [Rodgers *et al.* 2015]

but it is **difficult to quantify the cardinality** of sets or set intersections.

Linear diagrams are **easily scalable to more than three sets**, but lose their utility for Visual Analytics once a large number of sets are displayed. The difficult task for the viewer to visually process increasingly complex diagrams hinders efficient use of the visualization. Therefore, linear diagrams are not a suitable approach for the design of a set-based Visual Analytics software tool such as the Social Set Visualizer.

3.4.5 UpSet (2014)

Also in 2014, a significant contribution to set-theoretical visualizations was made through publication of the UpSet project [Lex *et al.* 2014].

UpSet is inspired by linear diagrams, and introduces an **intuitive, matrix-based representation** of set intersections that has been developed during earlier studies in the field of bioinformatics [Gratzl *et al.* 2013].

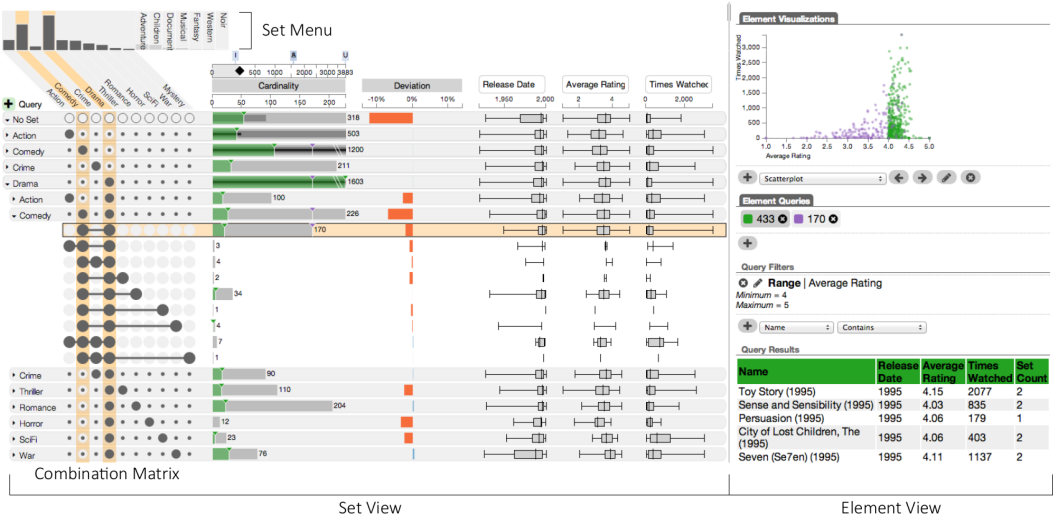


Figure 3.11: UpSet Combination Matrix-based Visualization [Lex *et al.* 2014]

Figure 3.11 illustrates the UpSet approach which is based on a combination matrix that encodes the different set intersections. The grid-based interface provides good oversight for the viewer and highlights the **future potential of UpSet for general-purpose Visual Analytics** of a large number of sets and set intersections. The user interface presented in the original publication is tailormade for statistical analysis in the field of bioinformatics and needs to be further simplified for general applicability.

3.4.6 UpSet-style Set Visualization in SoSeVi 2 (2016)

For the second version of the Social Set Visualizer, detailed in Publication III [Flesch *et al.* 2016] of this dissertation, an **UpSet-style visualization of sets** and set intersections was utilized.

Figure 3.12 illustrates the set-theoretical visualization in the second version of the Social Set Visualizer software tool. On the left hand side, all set intersections are encoded in a combination matrix. In the center of the visualization a bar chart depicts the cardinalities of the *Before*, *During*, and *After* time periods for each of the set intersections. On the right hand side, the migration patterns of social media actors between sets from subsequent periods, e.g. *Before* to *During* and *During* to *After*, are visualized. The UpSet approach was modified insofar that **logarithmic scales are utilized** in order to present both very large and very small sets and set intersections.

Through use of the UpSet-style approach, the visualization of sets in the second version of the Social Set Visualizer software tool is significantly improved, as **more than three sets can be visualized** while visual comparability is retained. This is an improvement of the first version of the Visual Analytics tool, in which only three sets could be visualized using the presented "exploding" Venn diagram technique.

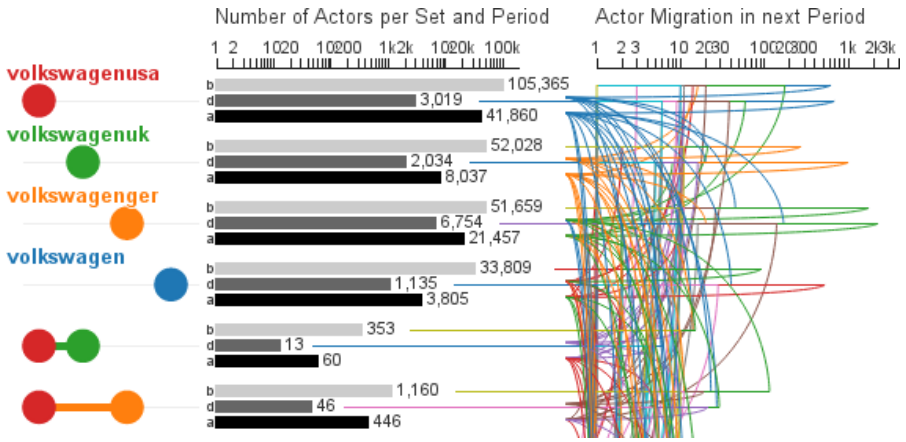


Figure 3.12: UpSet-style Set Visualization in SoSeVi 2 (Publication III [Flesch *et al.* 2016])

3.4.7 UpSetR (2017)

In 2017, with publication of the UpsetR visualization package, a **generalization of UpSet-style set visualizations** was presented to the research community [Conway *et al.* 2017]. UpSetR can be utilized to create static, non-interactive set visualizations.

Figure 3.13 displays an UpSetR-style visualization of six different sets. When compared to the original UpSet [Lex *et al.* 2014], it is noticeable that the combination matrix in the UpSetR visualization has been **rotated by ninety degrees**. Furthermore, the visualization has been significantly simplified and thereby streamlined.

It is immediately apparent that the UpSetR approach conveys a **large potential for Visual Analytics**, as the visualization of both set intersections and set cardinalities is very clear. Furthermore, the visualization is easily extendable and able to accommodate a large number of sets and set intersections.

The bar chart on top of the combination matrix is sorted in descending order, therefore allowing the most significant sets and set intersections to be shown at the top left of the visualization. On top of that, it provides **straightforward learnability for novice users** due to prior knowledge of how bar charts work. Once users comprehend the intuitive, matrix-based visualization of set intersections, the UpSet approach creates immediate utility.

Hence, overall **usability is significantly improved for visualizing a large number of sets** in comparsion to the other presented set-theoretical visualizations.

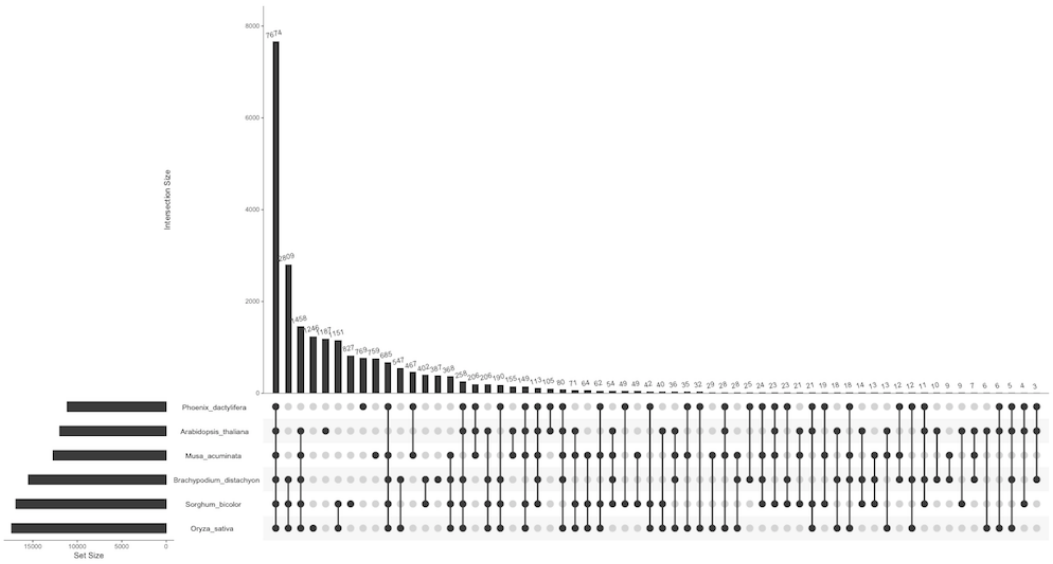


Figure 3.13: UpSetR Set Visualization [Conway *et al.* 2017]

3.4.8 UpSetR-style Set Visualization in SoSeVi 3 (2017)

In the third and final version of the Social Set Visualizer, UpSetR-style set visualizations are incorporated for the purpose of Visual Analytics of Big Social Data, as detailed in **Publication IV** [**Flesch** *et al.* 2017] of this thesis.

The UpSetR-style set visualizations in the Social Set Visualizer are utilized for **interactive set-based Visual Analytics of Big Social Data** along the two dimensions of time and space, as illustrated in Figure 3.14. Unlike the static visualizations in UpSetR [Conway *et al.* 2017], the Social Set Visualizer provides an interactive tool where the user can adapt the data that is displayed. Furthermore, it utilizes logarithmic scales and a color-coded bar chart in order to achieve a high level of interactive usability.

In contrast to the original UpSet [Lex *et al.* 2014], the Social Set Visualizer makes use of **server-side calculations** on its underlying Big Social Data corpus, and therefore is able to handle much larger volumes of data, with 100,000s to millions of actors included in the underlying sets. The Social Set Visualizer takes the original idea of UpSet much further than the client-side datasets which were presented in the original UpSet paper, and utilizes the novel set visualization technique to work with sets with large-scale cardinalities. As these large-scale set calculations are computationally intensive and require a lot of working memory, the analytical processing is offloaded to a specifically prepared backend system. This abstraction enables a smooth Visual Analytics user experience for users of the Social Set Visualizer. Moreover, the holistic approach of the Social Set Visualizer incorporates the majority of the Big Data Value Chain [Miller & Mork 2013] which was introduced earlier.

Most importantly, the visualization is adapted to handling and displaying large differences in set cardinalities by using **logarithmic scales** on both axes. Therefore, both small and large sets and set intersections are clearly identifiable and discernible. Even though the visualization is technically not area-proportional, the overall usability of the Visual Analytics software tool for a research context benefits

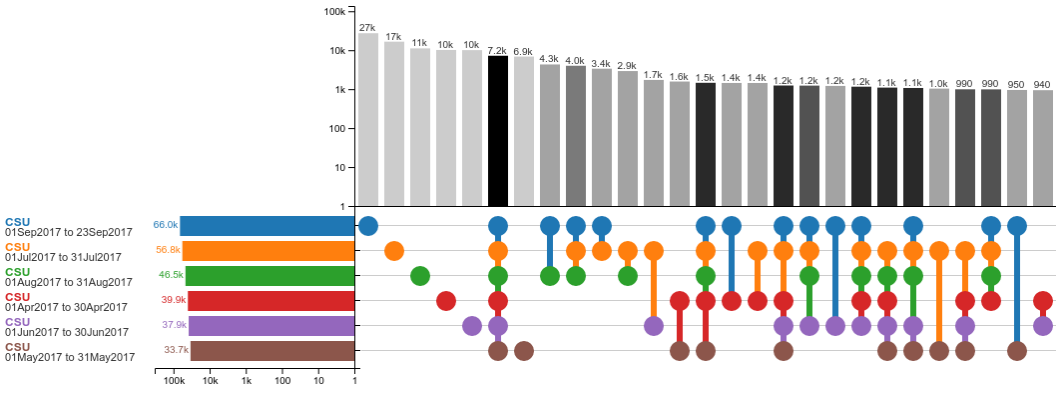


Figure 3.14: UpSetR-style Set Visualization in SoSeVi 3 (**Publication IV** [**Flesch** *et al.* 2017])

from this decision.

Furthermore, the UpSetR-style set visualization in the third version of the Social Set Visualizer makes extensive use of **color coding**. The horizontal bar chart is color-coded in **different shades of gray according to the number of sets that are intersected** with each other. This provides visual discernibility and comparability of set intersections, and allows the user to easily detect outliers in the dataset. The vertical bar chart is color-coded based on a categorical color scale, where **each base set is represented by one particular color** throughout the entire visualization. These colors are utilized throughout the intersection matrix, which increases usability of the overall visualization.

Both of these changes depict a **novel addition compared to both UpSet and UpsetR**, which only used one color for all sets in both the vertical and horizontal bar charts, and which provided no visual help through colors within the combination matrix.

3.5 Summary

In this chapter, an overview of the different approaches on set-theoretical visualizations were presented, as illustrated in Table 3.1. Based on the theoretical foundations, three essential design requirements for set-based visualizations in a Visual Analytics software tool such as the Social Set Visualizer have been elaborated, namelyy **interactivity, area-proportionality, and the number of sets that can be displayed**. Therefore, approaches performing strongly in relevant design requirements have been selected for implementation in the Social Set Visualizer.

Type	Year	Interactive	Area-proportional	Large number of sets	Application Domain
Euler diagram	1750	○	◔	◕	General Purpose
Venn diagram	1880	○	◑	◑	General Purpose
EulerAPE	2014	○	●	◑	General Purpose
Linear Diagrams	2014	○	◑	◕	General Purpose
UpSet	2014	●	◕	●	Bioinformatics
"Exploded" Venn	**2015**	●	●	◑	**Big Social Data Analytics**
UpSet-style SoSeVi	**2016**	●	●	●	**Big Social Data Analytics**
UpSetR	2017	○	●	●	General Purpose
UpSetR-style SoSeVi	**2017**	●	●	●	**Big Social Data Analytics**

Table 3.1: Comparative evaluation of set-theoretical visualizations

Thus, it was shown that the three versions of the Social Set Visualizer **contributed to the state of the art in set visualization** through design and presentation of the "Exploded" Venn Diagram (section 3.4.3), the UpSet-style (section 3.4.6), and the UpSetR-style set visualizations (section 3.4.8) in **Publication II** [**Flesch** *et al.* 2015a], **Publication III** [**Flesch** *et al.* 2016] and **Publication IV** [**Flesch** *et al.* 2017] respectively. Furthermore, the presented set-based visualizations improve both utility and usability of the set-based approach approach to Big Social Data Analytics that is presented in this thesis.

Chapter 4

Development

This chapter gives insight into the development of the Social Set Visualizer, which was created as an interactive Visual Analytics software tool based on Social Set Analysis methodology. First, the main objectives of the Social Set Visualizer software tool are highlighted. Henceforth, the underlying technological foundations of the Social Set Visualizer are presented, with particular focus on its approach to data storage and visualization. Furthermore, an overview is given on the software architecture of front- and backend. And, the different iterations on the IT artifact are introduced. Throughout this PhD project, three versions of the Social Set Visualizer software tool were implemented, which are detailed in **Publication II** [**Flesch** *et al.* 2015a], **Publication III** [**Flesch** *et al.* 2016], and **Publication IV** [**Flesch** *et al.* 2017] of this dissertation. Lastly, the deployment process will be outlined.

4.1 Development Objectives

Application performance, fault tolerance, and maintainability have been identified as the core objectives for software development [Spacey 2018] and are applied to the development of the Social Set Visualizer software tool to ensure a flawless development process.

Application Performance

With regard to the first key objective for the development of the Social Set Visualizer, **application performance** has been shown to increase the usability for users of a software tool [Etezadi-Amoli & Farhoomand 1996, Albert & Tullis 2013]. Good interactive performance can **positively influence the pace of insight generation** in Big Social Data Analytics. Even though this effect on the pace of insight generation is difficult to measure quantitatively, it becomes apparent through qualitative user feedback. This is why a holistic approach to application performance is adapted during development of the Social Set Visualizer, which focuses on both client- and server-side performance of the software tool.

In a web-based Visual Analytics dashboard such as the Social Set Visualizer, **client-side performance** is immediately apparent to the end user. However, client-side performance is difficult to measure and improve, as it mainly depends on the computational qualities of the utilized devices and general network connectivity [Viavant *et al.* 2002]. Consequently, the optimization of client-side performance is an essential theme during development of the Social Set Visualizer.

Moreover, **server-side performance** is important to process large volumes of Big Data. This creates computational difficulties when performing Big Social Data Analytics in an interactive, user-guided manner. From a software engineering perspective, server-side performance of web-based analytical applications can be improved through implementation of caching, parallelization, and pre-computation [Tilkov & Vinoski 2010, Subramanian *et al.* 1999]. However, the impact of these measures for the increase of server-side performance might be limited by the computational performance of the underlying hardware. The possibility of upgrading to more powerful hardware largely depends on the available budget.

Hence, a general decision problem concerning the optimal **distribution of computational tasks between client and server** emerges during development of the Social Set Visualizer. The UpSet software tool from the domain of bioinformatics, which was introduced in section 3.4.5, performs all computations in the client-side web browser. This client-side computation is feasible only as long as users of UpSet process small- to medium-size datasets that do not conflict with client-side limitations of a browser-based application, which include the maximum size of input files and the available working memory. Due to the multi-million datasets involved in Big Social Data Analytics, this approach is not feasible for the Social Set Visualizer. Therefore, most of the calculations are performed on the server side. By performing all computations involving raw social media data on the server-side, only aggregated results are transmitted to the client-side. Thereby, aspects of compliance and data privacy are strengthened in the Social Set Visualizer software tool.

Fault Tolerance

Fault tolerance depicts the second key objective for development of the Social Set Visualizer. It is based on the principles of **availability, reliability, correctness, and error handling**.

In this context, the term **availability** refers to the system's online status and whether it responds to the end user. Unavailability of the Social Set Visualizer software tool negatively affects the ability of researchers to utilize the tool and to generate insights through Big Social Data Analytics.

Reliability is directly linked to the concept of availability. It measures the mean time between failures of the application. Availability of the Social Set Visualizer software tool means researchers can navigate to the website and utilize the tool. It will be deemed unreliable in case the interaction with the user interface of the Visual Analytics tool fails with an error message. This example underlines the importance of reliability as an objective during development.

Correctness is a further component of fault tolerance. It aims for the Social Set Visualizer to provide the proper computational results to the user and to correctly implement the set-theoretical operations which are performed on the theoretical data model of Big Social Data. Due to the negative impact of software errors on the validity of scientific studies, the correctness of computational results remains an crucial objective, which can be mitigated through software testing or formal verification.

In addition, **error handling** as the final component of fault tolerance focuses on fail-safe behavior and graceful degradation.

Fail-safe behavior includes the preparation of continuity management in face of potentially catastrophic errors in the future. These errors can be mitigated through implementation of fail-safe measures such as database failovers and regular backups.

Henceforth, the concept of **graceful degradation** signifies that the software tool should gracefully reduce the quality of the user experience once an unforeseen event occurs, and thereby retain utility. This includes a responsive behavior to different errors, device types and screen sizes. Based on this definition, the Social Set Visualizer should be able to display both on a high-resolution 4K display used in a conference room and on a normal tablet. This responsiveness is needed to gracefully degrade the user experience depending on end-user device specifications, which in turn leads to a high level of fault tolerance in the resulting IT artifact.

Maintainability

Lastly, the objective of maintainability is set during development of the Social Set Visualizer software tool. It consists of two themes, namely **extensibility and interoperability**.

Extensibility concerns the question regarding the potential future development of the application. It addresses whether the underlying code base is sufficiently modular, meaning it allows future changes with only moderate effort and without unnecessary technical obstructions. Extensibility of an application is significantly influenced by the amount of **technical debt**. The concept of technical debt emphasizes that a badly engineered code base particularly slows future extensibility of an application, both in terms of feature development and removal of existing errors.

Furthermore, **interoperability** allows interfaces with other applications. As all data and computations within the Social Set Visualizer follow the theoretical model of Big Social Data, other tools such as the Social Data Analytics Tool are able to utilize datasets collected using the Social Set Visualizer and vice versa.

4.2 Technological Foundations

In this section, the technological foundations of the Social Set Visualizer in terms of the chosen data storage and visualization framework are presented. For both pillars, the underlying decision making process is detailed and relevant alternative options are discussed.

4.2.1 Choice of Data Storage

The evaluation of suitable data storage backends is vital for achieving good interactive performance for users of the Social Set Query Language with Big Social Data of various types and volumes. Implementation of the domain-specific Social Set Query

Language and its subsequent use within the Social Set Visualizer poses several challenges to the utilized database system. Therefore, a thorough **technical evaluation of various options for data storage** is required in order to identify a suitable database for the unique analytical workload of Social Set Analysis with large datasets. During the course of this PhD project, six different types of data storage systems which are well-suited for an implementation of the Social Set Query Language have been identified and evaluated:

- **NoSQL database systems (NoSQL)** such as **MongoDB**, a performant, cutting-edge approach to databases with a dynamic data schema based on the notion of documents with a flexible number of attributes, less strict compliance with atomicity, consistency, isolation und durability (ACID) than RDBMSs and mostly custom query languages.
- **Relational database management systems (RDBMS)** such as **PostgreSQL**, a solid, well-established type of database which provides strict compliance with ACID requirements on top of features such as a fixed data schema, stored procedures, transactions, and support of the Structured Query Language (SQL).
- **Distributed database systems** such as **Apache Spark**, a highly scalable, cutting-edge, distributed database with immense storage capacity, support of parallel analytical data processing, and cluster-based scaling approach.
- **Key-value database systems** such as **Redis**, a stable, in-memory database with native support for set intersections and a high level of performance.
- **Graph databases** such as **Neo4J**, a modern, enterprise-ready graph database implemented in Java with a custom SQL-style textual query language for graphs called Cypher paired with strict ACID compliance.
- **Implementation of a custom database system** using the **Go** programming language which is tailormade for the unique data schema of the Social Interaction Model and optimized for the OLAP-style workload of the Social Set Query Language.

The creation of an interactive Visual Analytics software tool has unique requirements on runtime optimization and cancellation of long-running queries. Therefore, a prototype data storage system supporting the Social Set Query Language has been implemented and evaluated for each of the six presented approaches.

NoSQL Database: MongoDB

MongoDB has been evaluated as storage solution for the Social Set Query Language based on its popularity as a NoSQL database. Due to its loose data schema definitions, Big Social Data can be stored in the same format as received from remote APIs. While the flexible data schema of MongoDB massively simplifies the import process

of historic Big Social Data from software tools such as SOGATO and SODATO, a NoSQL storage is not suitable for strict implementation of a data schema according to the the Social Interaction Model. In case social media data from different sources and collection timeframes is merged, a common data structure is vital. In a NoSQL database, it is hard to follow a strict schema definition, as any kind of input can be stored in any table. The evaluation of NoSQL data storage for the initial versions of the Social Set Visualizer as published in **Publication II** [**Flesch** *et al.* 2015a] has underlined this issue. On top of that, the tooling for MongoDB does not provide enough means for runtime optimization of database queries. Therefore, the initial version of the Social Set Visualizer was subsequently implemented based on a relational data schema which supports SQL. SQL lends itself to a highly dynamic way of building database queries, and provides many tools for performance improvements such as query optimization, table indexes, and stored procedures. MongoDB only provides custom querying functions which can be used through function calls in their NodeJS API. Paired with the difficulty of keeping Big Social Data in a certain structure and not-too-strict ACID compliance, NoSQL databases in general and MongoDB in particular do not depict a suitable option for data storage.

Relational Database: PostgreSQL

A relational database implementation of the Social Set Query Language was evaluated using the PostgreSQL database. PostgreSQL provides a strict database schema, which is defined in line with the theoretical model of Big Social Data, the Social Interaction Model. Furthermore, it gives mature tooling for query optimization purposes, including table indexes, query plan optimization and various configuration options for the PostgreSQL database service. During **Publication II** [**Flesch** *et al.* 2015a], it became apparent that the data import pipeline is more complex in a PostgreSQL scenario. Once the Big Social Data is loaded into the database system though, a lot of flexibility is gained through the availability of SQL and optimization features. Therefore, PostgreSQL was chosen as primary database backend for the Social Set Query language and it remains as such until the latest version 3 of the Social Set Visualizer presented in this dissertation.

Distributed Database: Apache Spark

After the implementation of the Social Set Query Language in **Publication II** [**Flesch** *et al.* 2015a] based on a PostgreSQL system, a thorough evaluation of potential performance improvements was performed. Hereby, special regard was set to scalable, distributed storage solutions such as Apache Spark for potential parallelization of analytical workloads. As PostgreSQL for data storage on its own has good, but not great performance, it was further attempted to empirically measure improvements in terms of execution time with our test data from other storage types.

The evaluation of the Social Set Query Language in Apache Spark resulted in mixed findings. On the one hand, the file-first approach of working with individual

files in a distributed storage system promises great potential when working with multi-gigabyte sized exports of Big Social Data such as social media data from Facebook. On the other hand, the parallelization of analytical queries showcased a significant performance lag between execution of the queries and the return of the results. Due to the strong performance requirements towards an interactive Visual Analytics dashboard such as the Social Set Visualizer, the database system on which the Social Set Query Language is implemented needs to return results very quickly.

Concluding from the evaluation, Apache Spark is not able to achieve better performance in terms of execution time due to long preparation time for parallel queries. Although the challenge of distributed storage of Big Social Data is well resolved by Apache Spark, low performance in terms of query execution time cannot be outweighted by the mentioned benefits. Therefore, the utilization of PostgreSQL RDBMS for our purpose should not be discontinued in favor of Apache Spark.

Key-Value Database: Redis

In Social Set Visualizer 2, the Social Set Query Language was implemented in Redis, a key-value database, and evaluated as the main database for set computations. This evaluation was described in **Publication III** [**Flesch** *et al.* 2016]. Redis provides **native set data types** which can be utilized with set-theoretical operations such as union and intersection. All set intersection calculations have been outsourced from PostgreSQL to Redis. A dedicated Redis instance now performs memory intensive set intersection calculations with a significantly better execution speed and pipes the calculation results back to the user-facing dashboard in real time.

Even though the **computational performance of set calculations** within Redis is outstanding, the severe limitations in terms of working memory and - by proxy - available research funds for extra-high-memory systems prevents a full focus on Redis as data storage solution. Any implementation of the Social Set Query Language with Redis as the sole data storage backend will be severely constrained by available working memory. Therefore, it is not financially feasible to implement all of the Social Set Query Language within Redis. Hence, in practice the use of Redis is always paired with a relational database such as PostgreSQL which depicts a back-up system to prevent loss of data.

Graph Database: Neo4J

Neo4J presents a modern graph-based database system. It was evaluated for storage of Big Social Data according to the Social Interaction Model and for its feasibility during evaluation of the third version of the Social Set Visualizer. It became apparent that the transformation of Big Social Data into a graph representation is a challenging undertaking. Even though Big Social Data structured according to the Social Interaction Model depicts several graph-like characteristics, the hierarchical overall structure could not be successfully stored in the graph-based database. The full potential of a graph-based database such as Neo4J could not be unleashed, as

Big Social Data does not depict a "true" graph with n-to-n connections, but rather many distinct trees which represent the individual interactions.

Furthermore, Neo4J provides a special graph-based query language that exhibits a very special optimization behavior. Thereby, dynamic formulation of database queries requires several optimization steps along the graph edges that are queried. Therefore, rapid iterations on the formulation of database queries as feasible in SQL is not possible in Neo4J. While the performance benefits of Neo4J with a proper graph-based dataset might provide more reassuring results, it did not hold for implementation of the Social Set Query Language. Hence, the decision against a graph-based database was made.

Custom Database Implementation

Lastly, an implementation of a custom Social Set Analysis database system in the Go programming language has been evaluated. Go provides excellent support for data structures. Thus, a conversion tool for Big Social Data to Go data structures in line with the Social Interaction Model was implemented for evaluation purposes.

Due to the size of the datasets involved in Social Set Analysis, various limitations in Go programming language were uncovered during implementation of a Social Set Query Language prototype. The respective bugs were reported to the project maintainers of Go[1] and resolved through a custom compilation of the Go programming language. Even though the implementation of a data storage backend was possible after these steps, the implementation of the actual query language posed as very cumbersome and time-consuming. Therefore, a decision was made against continuation of a custom database implementation in Go, as the extent of this task is beyond the scope of this PhD project.

4.2.2 Visualization Framework

The technology choice for realizing the dashboard visualizations is the D3.js [Bostock 2012] Javascript-based visualization framework which uses dynamic SVG images for data visualization. D3.js constitutes a lightweight and very extendable Javascript visualization framework. It uses cutting-edge web technology in order to facilitate the creation of complex visualizations. The flexibility provided by D3.js enables the creation of new kinds of interactive visualizations which are able to run on any device with decent processing resources. This includes Windows, MacOS and Linux based systems with screen dimensions up to 4K.

4.3 Software Architecture

The Social Set Visualizer dashboard consists of a client-side and a server-side component. Both are described in this section on software architecture.

[1] Issue #27635: "encoding/gob: tooBig is too small", full bug report available at https://github.com/golang/go/issues/27635

4.3.1 Frontend

All **client-side components** of the SoSeVi dashboard are implemented in the Javascript programming language which runs within the web browser of the user. The Visual Analytics dashboard is built on the React programming framework, which manages client-side state, user interface components, and routing within the single-page web application. Furthermore, several convenience frameworks for cross-browser utility functions and handling of dates are used. User-facing visualizations shown within the client-side dashboard are implemented using dynamic SVG graphics which are built upon the powerful D3.js visualization framework [Bostock 2012].

4.3.2 Backend

The backend of the Social Set Visualizer is split into two applications, namely **the data crawler and the API**. The data crawler handles data collection from Facebook in a dedicated process, which runs independently from the rest of the Social Set Visualizer. It queries the Facebook graph API and stores the results in a database. The API receives incoming queries which are formulated in the Social Set Query Language. The textual query is then interpreted and a list of sets which answer the query is created. For each set in this list, the API checks whether a cached version is available. If a cached version is available, the cached set is fetched and utilized for calculating the result of the query. If no cached version is available, the set definition is transformed into an SQL query which is sent to the database. The database returns all members of the queried set, which are in turn both stored in the cache and utilized to calculate the results that will be sent to the client-side user. In turn, the frontend utilizes this data to dynamically display the set visualizations.

4.4 Iterations on the IT Artifact

The development of the Social Set Visualizer follows an iterative process to enhance its performance and usability over time. In total, three distinct iterations of the Social Set Visualizer have been peer-reviewed and published during the course of this PhD project. In this section, all three iterations are presented and the contribution of each individual iteration is highlighted. Furthermore, the learnings from each iterative development stage are detailed.

4.4.1 First Version of SoSeVi (2015)

The first version of the Social Set Visualizer is presented in **Publication II** [Flesch *et al.* 2015a] of this dissertation. The software tool consists of five visualizations which are controlled by a central time selection brush. It provides a proof-of-concept interactive Visual Analytics dashboard for Big Social Data from Facebook, including several components inspired by Social Set Analysis methodology. Main feature of the Social Set Visualizer is the interactive selection of the *Event Window* and the *Before*

Figure 4.1: First iteration on the Social Set Visualizer dashboard

and *After* periods for analysis purposes. Figure 4.1 showcases the first version of the Social Set Visualizer.

In this first version of the Social Set Visualizer, an **"Exploded" Venn diagram** is chosen as set-theoretical visualization for the intersections of social media *Actors* from the *Before*, *During*, and *After* periods. This unique implementation of an Euler-based diagram was created to remedy shortcomings of classical Venn diagrams in terms of usability for the purpose of Visual Analytics. The novel "Exploded" Venn diagram is area-proportional and visually consistent in shapes and labels. It depicts a significant improvement that is detailed in section 3.4.3.

4.4.2 Second Version of SoSeVi (2016)

The second version of the Social Set Visualizer has been presented in **Publication III** [**Flesch** *et al.* 2016] of this dissertation. Moving towards the second version, the split of *Before*, *During*, and *After* periods is made more apparent in the user interface. Furthermore, the design settles on an UpSet-inspired visualization of sets, which is very suitable for the visualization of large-scale set intersections. The UpSet-inspired set visualization in the second version of the Social Set Visualizer is detailed in section 3.4.6. Application performance is significantly improved by optimizing the database layout and storage backend, as well as the utilization of advanced caching techniques based on a Redis in-memory database.

Figure 4.2 shows the second version of the Social Set Visualizer as an interactive tool for large-scale set intersection calculations. This version of the Social Set Visualizer depicts the first visual analytics tool to visualize migration flows in Big Social Data through means of clever set intersections along the two Social Set Analysis dimensions of time and space. Version 2 of the Social Set Visualizer project

as presented in this figure provides a significantly better interface for Social Set Analysis tasks of Big Social Data originating from Facebook. It depicts the first Visual Analytics tool to visualize migration flows between set intersections in Big Social Data. Furthermore, it allows an interactive selection of a relevant time period for analysis and calculation of intersections for *Before*, *During* and *After* periods over a big data-style dataset fetched from Facebook. Set intersections are dynamically encoded through a combination matrix. The migration of actors between time periods and set intersections is showcased through cardinality changes on the right hand side. User-provided event markers signify important real-world events with relevance for the underlying research domain of the investigating analyst.

The *DashboardView* (Figure 4.2) depicts the main view of the web application. It contains the main visualizations and is initially shown to the user. It consists of an overall activity visualization [F]. The researcher moves the time period selection tool [S] to navigate the data, and toggle data sets from different Facebook walls depending on the analysis tasks at hand. Based on the user-selected time period, which is labeled as the *During* period, *Before* and *After* time periods can be deducted

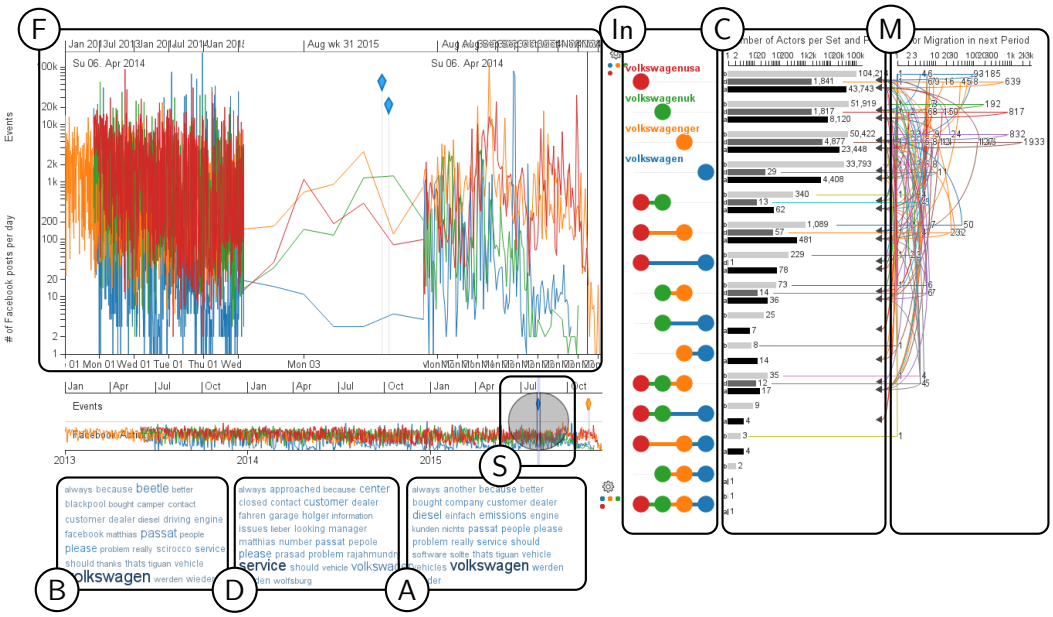


Figure 4.2: Version 2 of the Social Set Visualizer (SoSeVi) dashboard as of May 2016, showcasing 8M Facebook interactions from the *Volkswagen* pages adapted to the Social Set Analysis approach: [F] main activity chart zoomed in on the user-selected time period using the [S] selection tool. Underneath alphabetical word clouds for the time periods [B] before, [D] during, and [A] after are shown. On the right side, all set [In] intersections encoded as combination matrix, where [C] displays the cardinality of each intersection. On the very right, a visualization of period-over-period actor [M] migration between set intersections is displayed.

by looking at the beginning (earliest event) and the end (most recent event) of the underlying data. Three alphabetical word clouds underneath the main activity chart [F] illustrate the most important conversation topics in the [B]efore, [D]uring and [A]fter periods. This showcases the pluggable architecture of the Social Set Visualizer, which facilitates set-based content analysis tasks on the underlying dataset through use of the user-selected time frame.

To the right of the main activity visualization, the set intersection visualizations [In] are dynamically generated based on the user selection of the time period. Set intersections are encoded in a combination matrix. Each data source uses a distinct color. For each set intersection, up to three bar graphs are displayed in [C]. In case the set has a cardinality greater than zero, bar graphs for the *Before*, *During* and *After* periods are created. The bars are horizontally stacked, with the topmost bar signifying the *Before* period, the center bar *During*, and the lowest *After*. Right next to [C], a visualization of actor migration between periods and set intersections is displayed [M]. The Social Set Visualizer calculates all possible set intersections for each set with all sets of the following period based on the user-selected time period from the selection tool [S]. Therefore, the set intersections of all *Before* and *During* periods, but also of all *During* and *After* periods are calculated and shown. Migration flows can be individually analyzed by the user through interactive controls.

More information on [C] and [M] is detailed in Figure 4.3. Colored lines originating from the bar graphs in [C] are displayed in case the Social Set Visualizer has found set intersections with subsequent time periods. Thereby, the scale in [M] gives information on the cardinality of the migration between sets, and the curved line points to the destination set. Figure 4.4 illustrates how the researcher can **zoom in on a single migration path** and gather detailed information about sources and destinations of migrations. The cardinality numbers in [M] depict the actual migration volume, whereas the pointed arrows indicate the destination set of the migration.

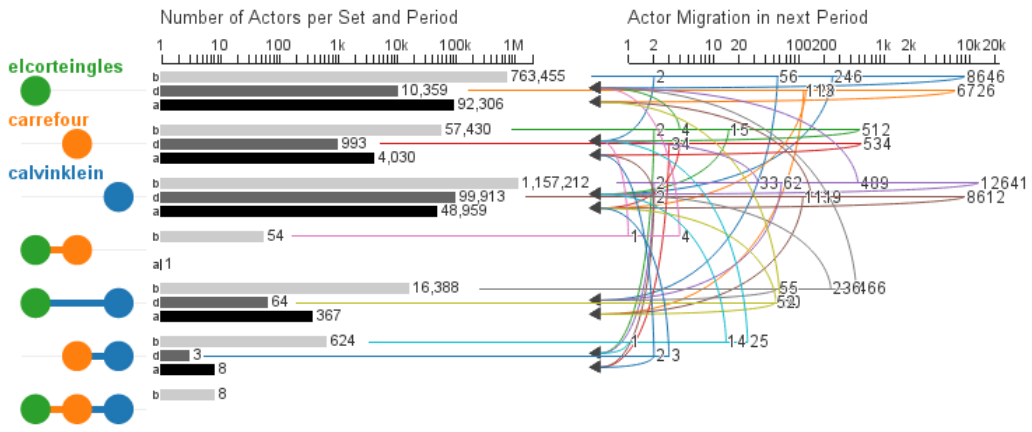


Figure 4.3: Visualization of actor migration over time and between set intersections

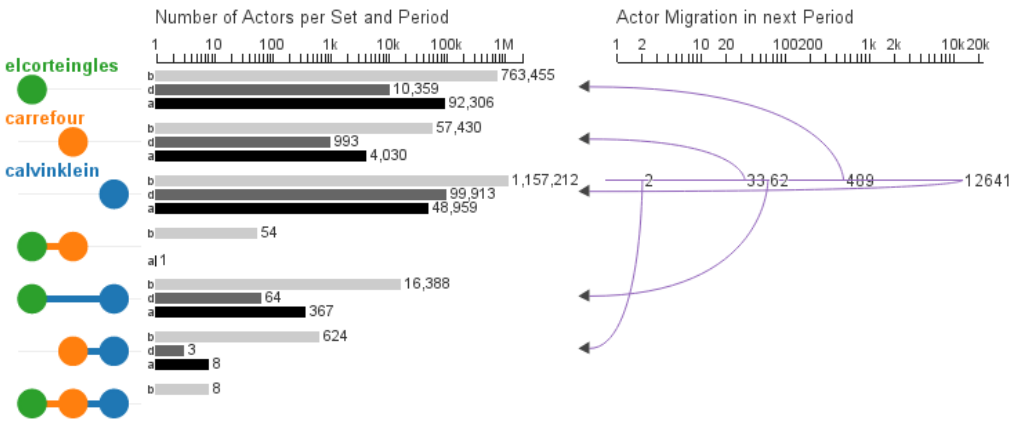


Figure 4.4: Visualization of actor migration originating from the *Calvin Klein* Facebook wall *Before* time period, showcasing strength and destinations of migration to set intersections in the *During* time period

This functionality implements the detail on demand principle which was introduced earlier.

RawdataView presents a detailed search interface for the underlying Facebook activity data. It is accessible to the user through various means by interacting with the visualizations of the *DashboardView*. It is a handy set of tools for analysis of actor mobility and cross-postings between different time periods and Facebook walls. *ActorsView* presents a dedicated interface for analysis tasks related to Actor Mobility across time and space of companies' facebook walls. The visualizations of actor mobility in *DashboardView* refer to *ActorsView* in order to provide the user with further details when requested.

In line with Action Design Research methodology, the main learnings from creation of version 2 of the Social Set Visualizer consist of the novel visualization of social media migration patterns and utilization of selection-matrix inspired set visualizations, an imitation of the Upset approach. Database and APIs have matured, but during development of the migration visualization and the optional on-mouseover detail-on-demand features, it became apparent that a domain-specific query language which allows for on-the-fly set definitions is very much needed. Hence, Social Set Query Language was introduced in the next version of the Social Set Visualizer for streamlining both development of the set-based visualizations and actual use in reproducible Social Set Analysis case studies.

4.4.3 Third Version of SoSeVi (2017)

The third version of the Social Set Visualizer shifted its full focus towards implementing Social Set Analysis methodology and combining it with an UpSet-inspired

Figure 4.5: Aggregate statistics on Facebook walls in SoSeVi 3

visualization of large-scale set intersections. Paired with the first implementation of the domain-specific Social Set Query Language, the third version of the IT artifact enables the user to dynamically prepare queries against the Big Social Data and to visualize the resulting set intersections. Social Set Visualizer 3 was first presented in **Publication IV** [**Flesch** *et al.* 2017] of this dissertation.

Figure 4.5 showcases the visual improvements in the listing of available data sources in Social Set Visualizer version 3. Each Facebook wall is aggregated so that emotions and conversation content is easily comprehensible. The improved user interface provides a much quicker way for identification of relevant data sources based on the displayed word clouds and reactions.

Figure 4.6 presents the UpSetR-style set visualization that has been implemented in the third version of the Social Set Visualizer. Visualizations are created based on textual queries in the Social Set Query Language, which can be either typed in by the user or is automatically generated using the data source and timeframe selection interface. The novel visualization allows for improved spatiotemporal analysis of Big Social Data, as the examination timeframes can be procedurally generated using the Social Set Query Language. Furthermore, extensive migration studies can be performed based on the newly introduced Social Set Query Language. This will be

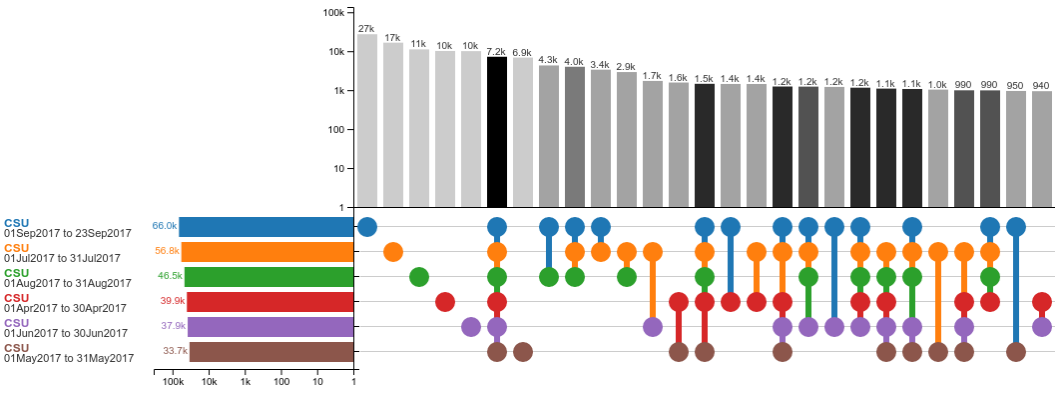


Figure 4.6: UpSetR-style set visualization of actor overlaps between political parties during the 2017 German federal election as shown in SoSeVi 3

further emphasized by the case studies in chapter 5.

Following Action Design Research methodology, the main learnings from creation of Social Set Visualizer 3 consist of the demonstration of feasibility for the newly created Social Set Query Language and various workflow improvements that result from the exclusive focus on the Upset-inspired visualization of large-scale set intersections. The built-in social media crawler has reached a good operational stability and presents a further learning on the value of vertical integration in a Big Social Data Analytics software tool. Thus, the Social Set Visualizer is able to cover as much of the Big Data Value Chain as possible. It has no further reliance on Facebook data gathered through external tools such as the Social Data Analytics Tool.

4.5 Deployment

All three versions of the Social Set Visualizer presented in this chapter have been deployed to **single machine with several CPUs and significant working memory**. Furthermore, no virtualization through tools such as VirtualBox, VMWare, or Docker has been used. The underlying PostgreSQL database runs on the same host as the Social Set Visualizer software tool with its two backend components, the data crawler and the API. Due to budgetary restrictions, no cloud-based deployment was made. Moreover, data privacy concerns are alleviated by the fact that the database that runs on the single machine is not exposed on the Internet. The Social Set Visualizer web application is protected by a username and password, which is shared with all researchers that utilize the tool.

4.6 Summary

In this chapter, the development of the Social Set Visualizer was detailed. It introduced the development objectives of application performance, fault tolerance, and maintainability that guide the development of the software tool. Furthermore, it described the technological foundations, and detailed the visualization framework. For the problem of data storage, a thorough evaluation of NoSQL, relational, distributed, key-value, graph-based, and custom databases was made, which resulted in the selection of PostgreSQL, a relational database, for the Social Set Visualizer. Then, software architecture on frontend and backend was outlined giving insight into the client-side and server-side components. After this, the three iterations on the IT artifact of this PhD project were introduced. Set-up and features of the first, second, and third version of the Social Set Visualizer are described in detail. Lastly, the deployment on a single machine without use of networking or virtualization was outlined.

Chapter 5

Evaluation

This chapter presents an extensive evaluation of the Social Set Visualizer. Multiple case studies have been performed on a wide variety of topics using large-scale Facebok datasets. In these, the Social Set Visualizer software tool has been utilized by other researchers and practitioners, and as a result contributed to **more than seven peer-reviewed publications** on the theme of descriptive and predictive analytics.

5.1 Descriptive Case Studies

The Social Set Visualizer software tool was evaluated through **four case studies in descriptive analytics** of Big Social Data. These case studies concern the Bangladesh factory disasters in the field of corporate social responsiblity, sports broadcasting for the two television stations TV2 Sport and NRK Sporten, Facebook activity across three music festivals Roskilde, Glastonbury, and Burningman, and lastly the emission scandal of Volkswagen.

5.1.1 Bangladesh Factory Disasters (2015)

The Social Set Visualizer case study on the topic of visualizing social media reactions upon specific events in the garment industry factory disasters in Bangladesh has been published in [**Flesch** *et al.* 2015b]. The study follows the Social Set Analysis approach and combines set-theoretical analysis of actors in Big Social Data with an Event Study Methodology in order to provide insights into large-scale events and their various facets.

In this case study, several research questions regarding the social media reactions to the Bangladesh factory disasters expressed on the Facebook walls of international clothing retailers were investigated.

The first version of the Social Set Visualizer was used for this case study. The user interface is illustrated in Figure 5.1. It allows selection of a time period either by using the date range fields in the top left, or by using the timeline selection tool which is marked as [2] in the graphic. Furthermore, selection of Facebook walls for analysis is possible through the color-coded input field at the very top of the screenshot.

Overall, the user interface contains six different visualizations which are annotated with their respective numbers:

(1) **Facebook activity chart.** It gives a detailed overview about the Facebook activity and real-world events in the selected timeframe. Based on the detail on demand

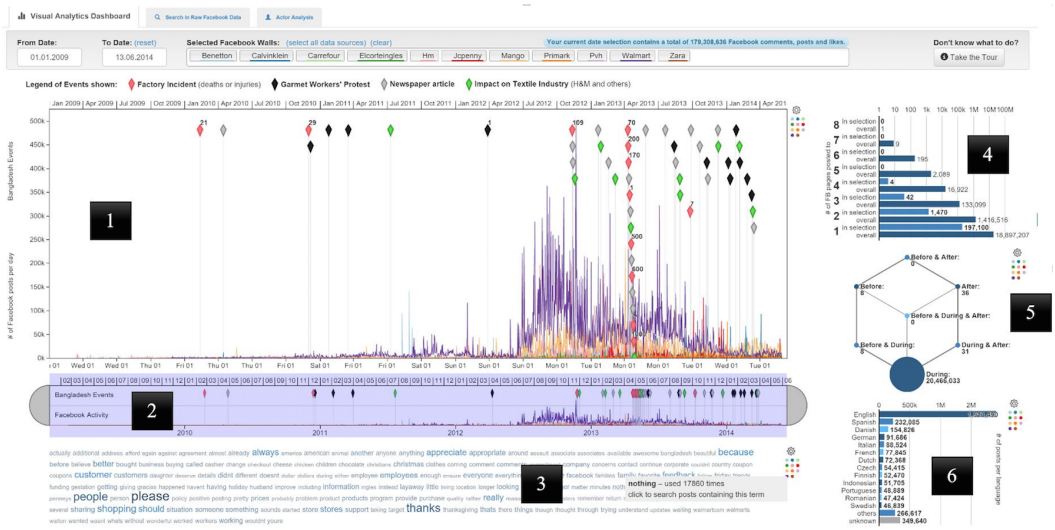


Figure 5.1: SoSeVi 1 used in case study on the Bangladesh factory disasters [Flesch *et al.* 2015b]

principle, different zoom levels show additional pieces of information. First, the aggregate number of data points for each Facebook wall is shown, then, each Facebook post is shown with the number of comments, and lastly, every individual post and comment is visualized.

(2) **Timeline selection tool.** It visualizes the timeline of real-world events on top of a line chart which displays the aggregate volume of Facebook activity. Its two brushes on the left and right side allow the user to select a timeframe by using drag and drop. Scrolling with the mouse wheel while hovering this visualization zooms in and out respectively.

(3) **Word Cloud.** An alphabetically sorted word cloud visualizes the top 200 most frequently occuring words from Facebook posts and comments. It is dynamically recalculated as the user interacts with the timeline selection tool (2) or changes the selection of Facebook walls.

(4) **Set intersections between Facebook walls.** This set-based visualization illustrates the mobility of social media actors between different Facebook walls. A set of all actors on each Facebook wall is calculated, then set intersections are visualized.

(5) **Set intersections between before, during, and after periods.** This set-based visualization illustrates the mobility of social media actors across the time dimensions. Sets and intersections for before, during, and after time periods are visualized via an "exploded" Venn diagram.

(6) **Language Distribution.** Based on natural language processing, a language classification of each post and comment within the selected time period is visualized.

For this case study, the Facebook walls of Bennetton, Calvin Klein, Carrefour, El Corte Ingles, H&M, JC Penny, Mango, Primark, PvH, Walmart, and Zara have been analyzed. Selected findings of this case study are:

(a) Global supply chain concerns with regard to Bangladesh garment factories have been expressed by Facebook users from as far back as 2009.

(b) There are many instances of authentic displays of support and expressions of empathy from Facebook users as well as robotic incidents of 'slacktivism'.

(c) Many of the uses of the word "please" were in relation to opening requests for new stores in the case of H&M.

(d) Protestors and activists employed different social media strategies on the different Facebook walls of companies but with little evidence for social influence (in terms of the number of likes and comments on their posts).

(e) Similarly, companies followed not only different corporate social responsibility strategies but also different social media strategies before, during and after the Bangladesh garment factory accidents.

(f) For almost all of the accidents, a majority of the users posting during the news-cycle do not return to the Facebook walls again. That is, social media engagement during factory accidents is detected as episodic and 'bursty' with little overlap to the "business-as-usual" period before or after the accident.

In this case study, the full archive of the social data from the Facebook walls of the 11 clothing retail companies was extracted using the Social Data Analytics Tool [Hussain & Vatrapu 2014b]. Given this data basis, I designed, developed and evaluated the Social Set Visualizer dashboard on this corpus of Big Social Data consisting of approximately 90 million data points. This case study analyzed social media data and presentated significant findings on the topic of corporate social responsibility. The set-based approach of the Social Set Visualizer software tool enabled a comparative, holistic study based on social media data from relevant companies in the clothing retail industry. Furthermore, set-based visualizations were utilized for visual analytics of various real-world events and social media reactions of an international audience.

5.1.2 Sports Broadcasting by TV2 and NRK (2016)

A comparative case study between two major Scandinavian sports broadcasters was carried out using the Social Set Visualizer. In this case study the Facebook pages of TV2 Sporten, a Danish TV station, and NRK Sport, a Norwegian TV station, were analyzed to give insight into differences in terms of audience reactions to important

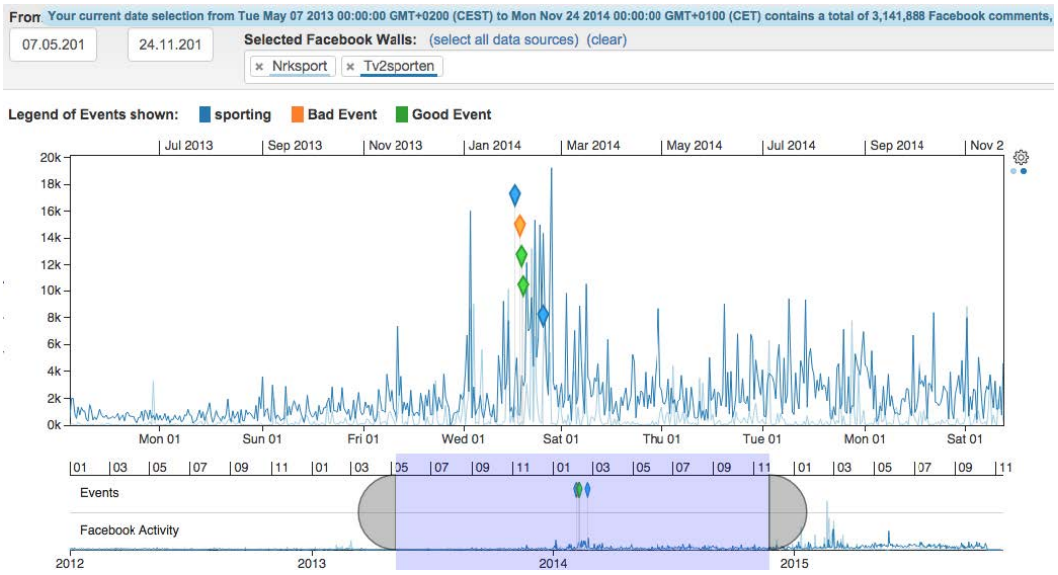


Figure 5.2: Temporal distribution of total Facebook activities for NRK Sport and TV2 Sporten (SoSeVi 1) [Hennig *et al.* 2016]

sports events. The case study has been published as a student project in the Big Data Analytics course, to which the Social Set Visualizer tool was provided [Hennig *et al.* 2016].

Figure 5.2 showcases the temporal distribution of Facebook activity for both TV stations in the Social Set Visualizer dashboard. The visualization enables researchers to identify event-related spikes, seasonal variations, and demographic patterns in the data at hand. Furthermore, it provides proof for large-scale temporal patterns, as data volume is steadily increasing over the two-year observation window of the study. Subsequently, the comparison of activity numbers during live broadcasting and re-run programs showed significantly more Facebook activity during live broadcasting.

Figure 5.3 showcases the utilization of Social Set Visualizer to illustrate overlaps between the Facebook pages of TV2 Sporten and NRK Sport. Additional findings of this case study include the quantification of sentiment distribution between "good" and "bad" sports events. Furthermore, the researchers were able to discover that the distribution of emotional patterns expressed through Facebook reactions changed over time. The development of positive, negative and neutral sentiment activity on the Facebook pages during the olympic games was analyzed using the Social Set Visualizer. It was found that TV ratings are significantly affected by sports and gender, as "single events showing male biathlon athletes have more viewers, but female biathlon drew more viewers than male biathlon in total".

In this case study, the first version of the Social Set Visualizer software tool was used for Visual Analytics of Big Social Data on a dataset of 7,532,000 pieces of Facebook activity. Its set-based approach to Visual Analytics enabled the student

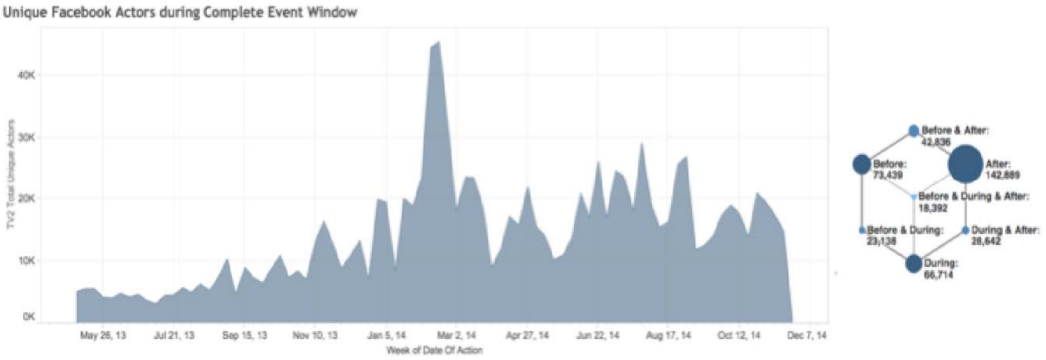


Figure 5.3: Unique Facebook actors during complete event window on TV2 Sporten (SoSeVi 1) [Hennig *et al.* 2016]

researchers to quantify overlaps in social media activity. This was done by calculating and visualizing set intersections of the before, during and after time periods of real-world, TV-broadcasted sports events.

With regard to the research question of this thesis, this case study shows that the utilization of a set-based approach to Big Social Data Analytics generates unique and relevant insights. Furthermore, student researchers were quickly able to use the Social Set Visualizer software tool to produce research findings that were presented to an international audience at the 2016 IEEE Conference on Big Data.

5.1.3 Roskilde, Glastonbury & Burningman Festivals (2016)

A comparative case study using the Social Set Visualizer was conducted on the topic of festival analytics. In this study, the Facebook activity of three world-famous music festivals, namely Roskilde Festival (Denmark), Glastonbury (UK), and Burningman (USA) is analyzed using our set-based approach to Big Social Data Analytics.

This case study is attached as **Publication III** [**Flesch** *et al.* 2016] to this dissertation. It was created using the second version of the Social Set Visualizer software tool. Methodology and findings were presented to an international audience of set visualization experts at the Set Visualization and Reasoning workshop (SetVR) at Diagrams conference 2016.

The case study showcases that all three festivals have interactions with several hundreds of thousands actors on Facebook. Figure 5.4 illustrates the set intersections that were performed with the Social Set Visualizer on this dataset. The visualization shows cardinalities of various large-scale sets with between 200 and 500,000 actors each. For each festival, sets for the user-selected before, during and after periods are calculated. Then, pairwise intersections between all sets are performed, and their cardinality is visualized. The set intersections between the three festivals Roskilde, Glastonbury and Burningman emphasize the fact that Glastonbury and Roskilde, but also Glastonbury and Burningman share a significant amount of users. However,

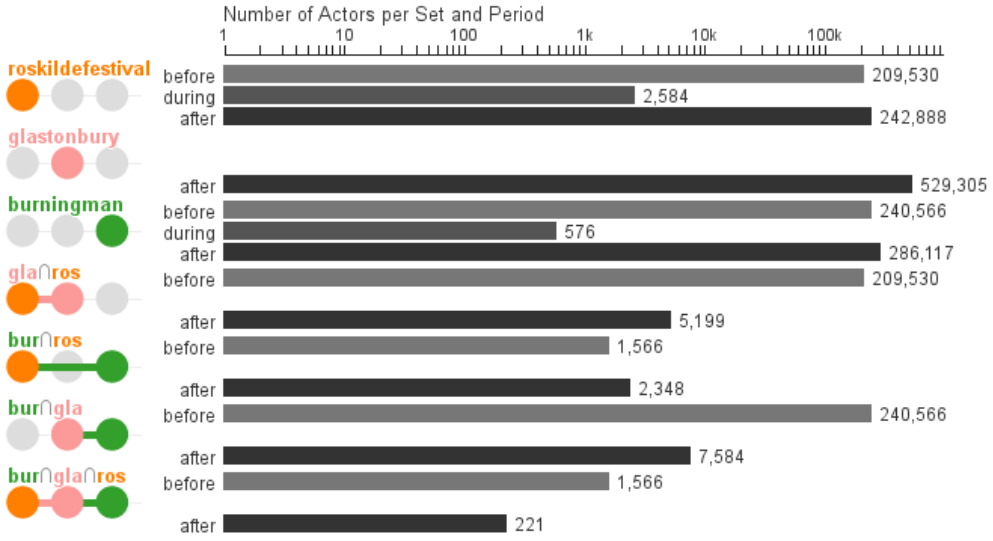


Figure 5.4: Visualization of set intersections and set intersection cardinality before, during, and after the user-selected time period, illustrating the distribution of social media actors over time and space. **Publication III** [**Flesch** *et al.* 2016]

the intersection between Burningman and Roskilde is very small. Furthermore, the number of users who are active on all three festival pages during the examination period is very low, only around one thousand users.

In this evaluational case study, the Social Set Visualizer software tool has used data collected from three different Facebook walls with massive user bases, totaling more than 10M data points. Therefore, this case study demonstrates the feasibility of large-scale set intersections in the interactive visual analytics dashboard based on dynamic user-driven selection of investigation timeframes, and thus presenting an example case for answering the first research question.

5.1.4 Volkswagen Emission Scandal (2016)

The Social Set Visualizer was applied to the Facebook walls of Volkswagen in an unpublished case study on the Volkswagen emission scandal. Figure 5.5 depicts a screenshot of the set-based visualizations in this case study.

The capabilities of Social Set Analysis are showcased through a visualization of set intersections between the different Volkswagen entities and a visualization of migration patterns between each set intersection. The Social Set Analysis methodology was demonstrated by comparisons of actor movements across dimensions of time and space in order to visualize social media migration patterns.

The findings of this study emphasize the fact that dominant discussion topics can quickly change when a scandal emerges from news reports. Once reports on the emission scandal were published, the user activity on the Facebook walls of four

corporate entities of Volkswagen displayed very similar patterns.

In this case study, it was shown that the Social Set Visualizer software tool is able to perform large-scale set-based Big Social Data Analytics in order to compare different country pages of the Volkswagen AG. The dataset consisting of 8M Facebook interactions from four corporate entities in three countries, namely UK, USA and Germany, outlines the utility of the Social Set Visualizer in comparative event studies of Big Social Data.

With regard to the first research question of this thesis, the second version of the Social Set Visualizer as utilized in this case study implements a novel approach to set-based visualizations inspired by UpSet [Lex *et al.* 2014]. Through this, the visualization quantifies social media migration patterns which happen before, during and after the user-selected time periods in relation to important real-world events. This unique approach to Big Social Data Analytics allows an efficient comparison of four different social media presences of Volkswagen AG in light of the emission scandal.

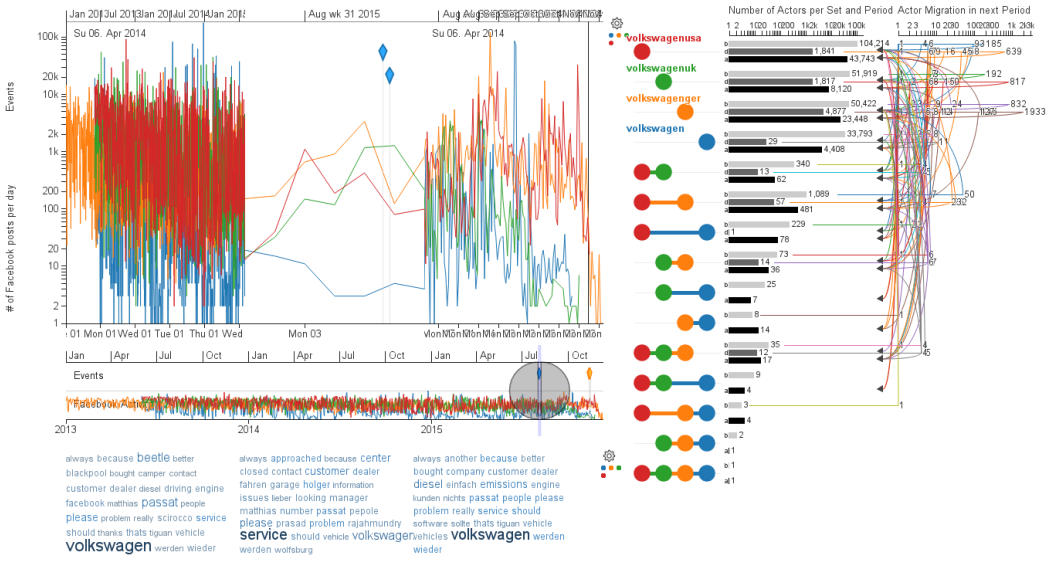


Figure 5.5: SoSeVi 2 displaying 8M interactions from the *Volkswagen AG* Facebook pages in a study on the emission scandal. [**Flesch** *et al.* 2016]

5.2 Predictive Case Studies

Furthermore, the Social Set Visualizer was utilized in three predictive case studies concerning audience prediction for Roskilde music festival, Facebook-based sales forecasting for Nike, and voting prediction of the German federal election.

5.2.1 Sales Forecasting for Nike (2016)

A case study on the topic of sales forecasting using social media data by example of the Nike corporation was co-authored with several students from the Big Data Analytics course [Boldt *et al.* 2016]. Like the previous case study, this case study was presented at the 2016 IEEE Conference on Big Data in Santa Clara, USA. The data was fetched using the Social Data Analytics Tool [Hussain & Vatrapu 2014a]. With this data, the Social Set Visualizer software tool was initially used to conduct an event study based on data from the Nike Facebook wall, in which the time windows of different events were compared and events of interest were surveyed.

The findings of this study suggest that Facebook data contains informational value and that simple regression models provide a high forecasting accuracy for prediction of future sales for Nike. In this paper, we describe the exact utilization of the Social Set Visualizer as follows:

> *"In order to understand whether the events actually drive the abnormal activity, we used the Social Set Visualiser (SoSeVi) tool and a netnographic study on the relevant Facebook pages. By applying the SoSeVi tool we are able to conduct simple text analytics that gives a better understanding of the content that drives the actual spike(s)."*

Therefore, through interactive definition of sets in the Social Set Visualizer, the authors were able to select and compare Facebook activity from the before, during, and after time periods of the events of interest. Further findings of the event study performed with the Social Set Visualizer are described as follows:

> *"From the aggregated event study, we identified that some types of events had more traction than others. Campaigns with hashtags attained to it had more traction on Facebook on average, compared to events such as product launch or negative media. This is in line with recommendations from a social media-consulting firm, which explain the positive impact of hashtags in engagement in social media platforms."*

In this evaluational case study, the Social Set Visualizer was utilized for an investigation of the Nike Facebook page. Several findings were documented, and both long- and short-term developments in the Big Social Data at hand were identified. The student co-authors of this paper highlight the ease of use of the Social Set Visualizer software tool. Furthermore, they were able to quantitatively compare events with positive and negative overall sentiment based on a set-based approach.

5.2.2 Roskilde Festival Artist Audience Overlaps (2017)

The predictive case study on the Roskilde festival stems from an unpublished research project in which our research group volunteered to perform data analytics at the 2017 Roskilde Festival. For this case study, the Facebook walls of several hundred artists that were scheduled to play at Roskilde Festival 2017 were both fetched and analyzed through use of the Social Set Visualizer.

The Roskilde Festival case study has two major goals. First, it is to investigate if there is a way to optimize the scheduling of artists on different stages with a special focus on crowd safety. This is needed, because critical crowd safety situations can potentially arise as large crowds move from one artist to the next one which plays at a different stage. The goal is to investigate if these critical crowd safety situations can be **predicted by crowd movements based on social media data** from Facebook.

The second goal is to **predict the largest concert at the festival**. For this, the Social Set Visualizer software tool is used to quantify audience overlaps between the Facebook audience of Roskilde Festival itself and each of the scheduled artists. The cardinality of each set intersection is quantified, and a prediction of the TOP 10 artists is presented.

Prediction of Crowd Movement

With regard to the first objective, the Social Set Visualizer enables prediction of crowd movement. Thus, actionable findings with special relevance for crowd safety efforts at Roskilde Festival have been generated through use of the Social Set Visualizer software tool.

The methodology is based on information of artists who are scheduled in sequence to each other on separate stages. They are seen as potential source of massive crowd movement between concert stages and therefore as a potential threat and hazard to crowd safety. For this, we apply the Social Set Analysis methodology to the Facebook data fetched through the built-in crawler of the Social Set Visualizer. Hence, we can quantify shared audiences between scheduled artists and highlight potential crowd movement safety risks well in advance. For this purpose, the artist Facebook audiences were intersected first with the Roskilde Festival Facebook wall, in order to identify fans of each artist who are likely to go to Roskilde Festival 2017. Subsequently, these artist and Roskilde Festival set intersections were again intersected with each other with a special focus on artists with neighboring time slots.

The cardinalities of these inter-artist set intersections are calculated and visualized with the Social Set Visualizer software tool. Afterwards, the calculated cardinality information can be overlayed on top of the festival programme. Figure 5.6 showcases the scheduled concerts and the cardinality of each pairwise intersection for the festival on Friday, 30 June 2017. This case study depicts the **first time crowd movements in a music festival are predicted through the use of Big Social Data Analytics**.

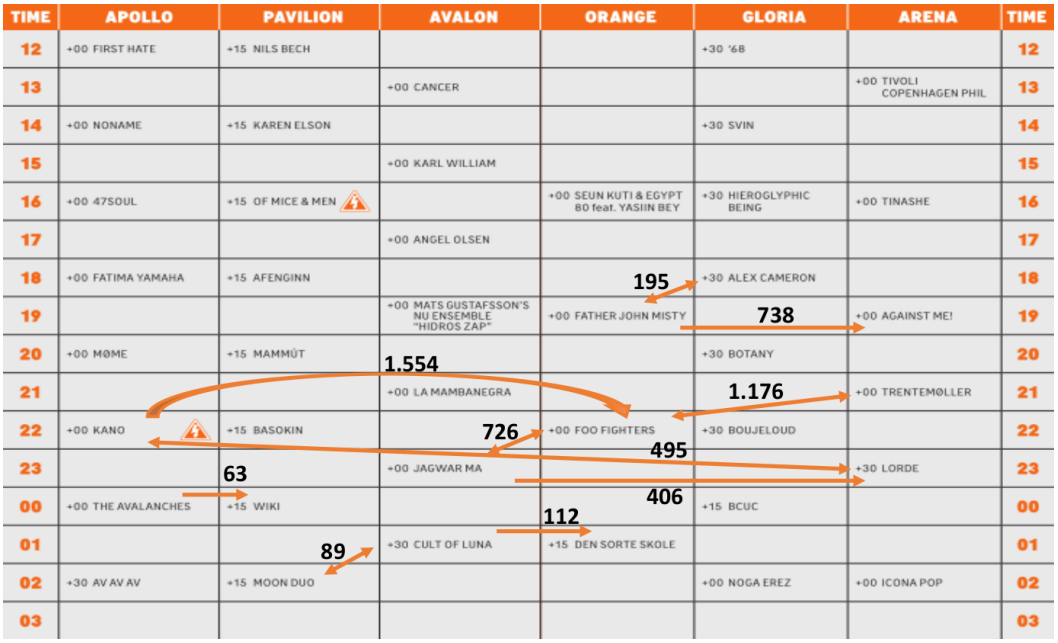


Figure 5.6: Overlaps between Facebook audiences of different artists at Roskilde Festival 2017

The detailed predictions for potentially safety-critical large-scale crowd movements during the 2017 Roskilde Festival were as follows:

(1) For **Thursday, 29 June 2017**, the concert of Elza Soares at Avalon stage with Solange immediately following at Arena stage was identified as highest risk for massive crowd movement of 2,485 festivalgoers between stages based on Big Social Data.

(2) For **Friday, 30 June 2017**, the concert of Kano at Apollo stage with Foo Fighers at the same time at Orange stage was identified as highest risk for massive crowd movement of 1,554 festivalgoers.

(3) For **Saturday, 1 July 2017**, the concert of Anthrax at Arena stage with Arcade Fire immediately following at Orange stage was identified as highest risk for massive crowd movement of 5,111 festivalgoers. Further movements from one stage to the other are expected in case festivalgoers decide that the Arcade Fire concert will be more entertaining.

Due to a lack of data from the festival organizer, an exhaustive validation of the stated predictions is not possible. Official audience measurements are required in order to perform a thorough comparison of the prediction with actual numbers. This predictive case study focuses on the utility of the Social Set Visualizer in a festival analytics scenario. It highlights that the set-based approach to Big Data Analytics is applicable to a variety of real-world data science research questions.

Prediction of Concert Attendance

The second goal of this case study is to predict concert attendance based on social media data from Facebook. This prediction uses the simple heuristic of pairwise intersections between the Roskilde Festival Facebook page and each artist. The cardinality of these set interactions is calculated in order to identify the artists which are most popular with the Facebook audience of Roskilde Festival.

The underlying assumption of this approach to predicting concert attendance is that the social media audience on the Roskilde Festival Facebook page is representative of the actual on-site audience of festivalgoers. If this assumption holds, the social media interactions with both the Facebook pages of the artists and of Roskilde Festival carry some sort of signal that lets us predict a relative ranking of how many people will show up at the actual concert during the festival.

Figure 5.7 showcases the prediction of concert attendance at Roskilde Festival 2017 based on set-based artist overlaps with the official Roskilde Festival Facebook

RF 2017 Concert Attendance Prediction via Social Set Analysis of Facebook Audience

27 Jun 2017 by Benjamin

#	Artist	# FB overlap w/ Roskilde
1	Foo Fighters	3,269
2	Arcade Fire	2,053
3	Phlake	2,046
4	Karl William	1,738
5	Emil Stabil	1,734
6	The Hellacopters	1,719
7	Sort Sol	1,574
8	The Weeknd	1,349
9	Royal Bood	1,226
10	Anthrax	1,103
11	Justice	1,076
12	Father John Misty	1,050
13	Trentemöller	956
14	The XX	946
15	Carl Emil Petersen	919

Figure 5.7: Prediction of concert attendance at Roskilde Festival 2017 through set-based artist overlaps with Roskilde Festival Facebook page

page. The most popular artist according to its Roskilde Facebook audience overlap is Foo Fighters with 3,269 users. Arcade Fire and Phlake are second and third most popular artists.

Several hundred artist Facebook walls have been fetched and compared for this case study. In line with Gartner's three types of descriptive, predictive, and prescriptive data analytics which were presented in the introductory chapter, this case study not only performs descriptive analytics but also shows predictive character.

Actionable insights into crowd safety operations were presented to the festival management and preparations were made for larger-than-expected crowd movement between stages. Furthermore, concert attendance was predicted and it was forecasted that Foo Fighters and Arcade Fire will draw the largest audiences at the festival. This prediction turned out not the be correct, as The Weeknd, an artist that I ranked 8th in my prediction, drew the largest audience of the festival. A further validation of these findings is only tangentially related to the core of this thesis, and has therefore been postponed as future work.

This case study contributes to the first research question of this thesis by providing an example use case for predictive analytics using the Social Set Visualizer. The software tool and its set-based approach have been utilized to generate meaningful insights in the festival analytics case, hence their predictive utility in real-world scenarios is showcased.

5.2.3 German Federal Election (2017)

A further extensive predictive case study on the 2017 German federal election has been performed using the Social Set Visualizer. It is attached as **Publication IV** [**Flesch** *et al.* 2017] to this dissertation.

This case study concerns all major German political parties, namely SPD (socialists), CDU (conservative), CSU (Bavarian party), FDP (liberal), AfD (far right), Bündnis '90 Die Grünen (green party), and Die Linke (far left) parties. Using the Social Set Visualizer, the Facebook walls of each political party have been fetched for the time from January until September 2017, the month of the federal election. The dataset used in this case study contains more than 15M data points with a total of 1M unique actors.

Figure 5.8 depicts the Social Set Visualizer selection interface in which the user can select Facebook walls and timeframes for analysis. It showcases a selection of all seven political parties for the time from January to September 2017. The sparkline-style activity visualization included in the figure hints at the fact that overall activity of all parties peaks in September 2017, the month of the federal election.

The Social Set Visualizer enables analysis of audience overlaps between German political parties during the 2017 election, which is showcased in Figure 5.9. The visualization follows the UpSet approach on visualization of large-scale set intersections.

From the dataset at hand, it becomes apparent that far-right AfD party has the highest number of Facebook members across all parties, a total of 295,000 individuals. SPD depicts the second-largest party, with a total of 221,000 individuals.

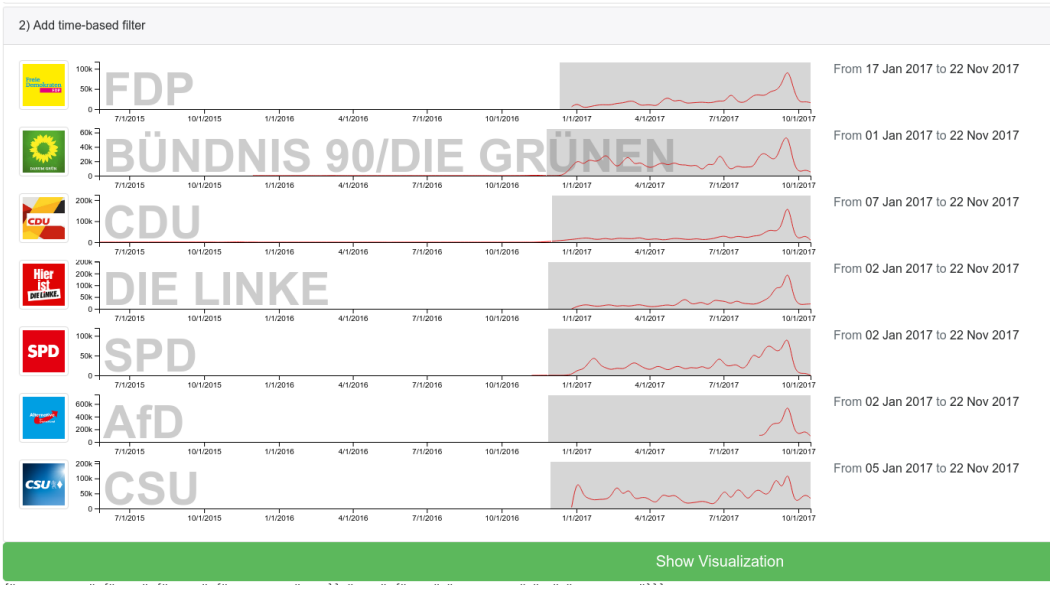


Figure 5.8: SoSeVi 3 Facebook wall selection interface (**Publication IV** [Flesch *et al.* 2017])

Furthermore, pairwise overlaps between political parties follow the parties' closeness and positions on the political spectrum:

1. We observe that more than 27,000 individuals were active both on the Bavarian *CSU* and the far-right *AfD* Facebook pages, displaying the biggest audience overlap between two political parties.

2. The second major audience overlap is between far-right *AfD* and far-left party *Die LINKE* with 9,600 individuals.

3. The third largest overlap is between Bavarian *CSU* party and liberal *FDP* party with more than 9,500 individuals active on both parties' Facebook pages.

4. Fourth largest overlap is between social democrats *SPD* and far-left *Die Linke* with *9,500* individuals, followed by fifth largest overlap between *SPD* and the green party *Bündnis '90 Die Grünen* with 9,100 individuals active on both Facebook pages.

5. Angela Merkel's conservative *CDU* and her Bavarian sister party *CSU* depict the sixth largest overlap with 8,700 individuals.

Further overlaps between political party Facebook audiences are visualized in the figure. The major overlaps identified seem to follow the parties' closeness on the political spectrum, even though at the moment, we cannot explain the detailed reason for the relative differences in cardinality between overlaps such as *CSU*|*AFD* and *SPD*|*Die Linke*.

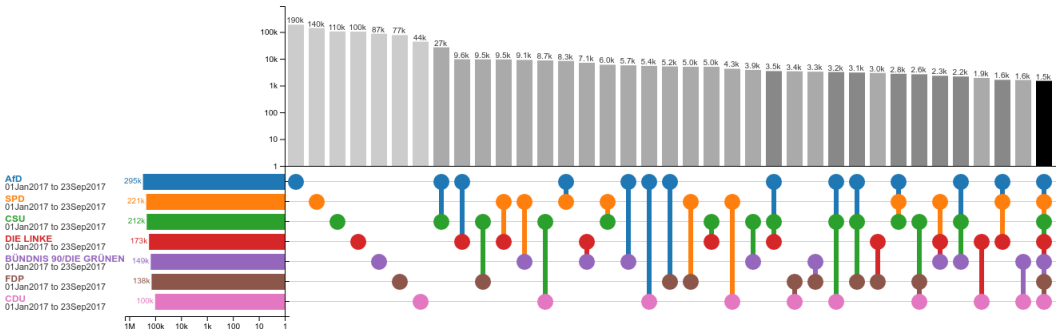


Figure 5.9: Audience overlaps between political parties during the 2017 German federal election visualized in SoSeVi 3 (**Publication IV** [**Flesch** *et al.* 2017])

Figure 5.10 quantifies the growth patterns of each party over the period leading up to the election. From this, we can observe the following:

1. For all parties, the final month of campaigning, September, was the best month in terms of total number of individuals they interacted with.

2. No party showcases a steady, consistent growth story. All of them have at least one month where they actually decreased their audience compared to the previous month.

3. Comparing the compound monthly growth rate (CMGR), both leftist *LINKE* (+35.9%) and Angela Merkel's party *CDU* (+31.7%) depict the biggest growth over the whole period of investigation. Both are also the only two parties that show their weakest overall activity in April and their strongest activity in September.

4. With only a 10.6% growth rate over the whole election campaign, the Bavarian *CSU* party had the lowest growth in audience of all seven parties.

5. All other parties *SPD*, *FDP*, *Bündnis '90 Die Grünen*, and *AfD* express a growth rate of roughly 20% per month.

	April	May		June		July		August		September		All months	
Party	#	% chg	#	% chg	#	% chg	#	% chg	#	% chg	#	Sparkline	CMGR
AfD	63.1k	-8%	58.4k	8%	63.4k	-6%	60.0k	35%	92.8k	40%	155.0k		19.7%
CDU	11.8k	16%	14.1k	18%	17.2k	32%	25.4k	-11%	22.8k	51%	46.7k		31.7%
CSU	39.9k	-18%	33.7k	11%	37.9k	33%	56.8k	-22%	46.5k	30%	66.0k		10.6%
FDP	25.1k	6%	26.8k	-25%	21.5k	34%	32.5k	29%	45.8k	25%	61.1k		19.5%
GREEN	18.6k	11%	20.9k	17%	25.1k	-47%	17.1k	58%	40.4k	10%	44.9k		19.3%
LINKE	18.7k	52%	39.3k	0%	39.3k	-2%	38.5k	8%	41.8k	52%	86.7k		35.9%
SPD	31.4k	-18%	26.7k	27%	36.8k	-10%	33.5k	54%	72.5k	13%	83.8k		21.7%

Figure 5.10: Political party growth rates during 2017 German federal election (**Publication IV** [**Flesch** *et al.* 2017])

6. In August, penultimate month of the 2017 election campaign, current chancellor Angela Merkel's parties *CDU* and *CSU* both decreased in the number of individuals that interacted with their Facebook pages by a total of 69.3k people (-11% and -22% respectively). This is particularly interesting because one would expect that during August, at the peak of campaigning, both sister parties would continue to push very hard. This decrease could be explained with summer holidays for the shared campaigning team.

7. Also in August, socialist *SPD*, as biggest rival of *CDU*|*CSU*, grew their audience at 54%. With a total of 72.5k individuals, *SPD* reached a larger audience on Facebook than both *CDU* (22.8k) and CSU (46.5k) combined.

The original publication attached to this dissertation contains many more findings on this topic. In this evaluational case study, I have shown that the set-based approach of the Social Set Visualizer can be applied to a multitude of research questions in election analytics. In this large-scale election case study, nearly 1 million social media actors and their interactions with German political parties were aggregated and quantified, thus providing a unique perspective on an election with international significance.

Both descriptive and predictive analytics have been presented in this case study. Even though the final election outcome was not correctly predicted by the relative shares of unique Facebook audience members, it is still a significant finding that the German federal election can not be predicted based on Facebook data. This might be caused by methodological limitations of this case study, which only analyzed the national party Facebook pages, and no regional or communal Facebook pages. However, as many parties such as election winner CDU do not have significant activity on their Facebook page, it is an open question whether Facebook has a significant signal for prediction of parliamentary elections due to special circumstances in German demographics.

This case study has contributed to the first research question of my thesis by generating a multitude of findings in an election analytics scenario. It showcased the utility and flexibility of the final version of the Social Set Visualizer software tool and its contribution to state-of-the-art research on important societal topics. The software tool downloaded a large-scale Facebook dataset, performed large-scale set intersections, and visualized the results for analysis. Therefore, the Social Set Visualizer implemented the entire Big Data Value Chain in this case study.

5.3 Summary

In this chapter, the Social Set Visualizer software tool was thoroughly evaluated over the course of seven case studies. In four of these case studies, the tool was utilized for purpose of descriptive analytics, while in the three remaining case studies the Visual Analytics tool was applied for predictive analytics.

In descriptive analytics of the Bangladesh factory disasters case study, it was shown that the first version of the Social Set Visualizer can generate novel research findings regarding a global phenomenon that impacts eleven large-scale retail companies on social media. In the comparative study between two sports broadcasters, the software tool was used to calculate overlaps and visualize actors before, during, and after windows. A three-way intersection between concertgoers and a quantification of festival churn rates was performed in the analysis of Roskilde, Glastonbury, and Burningman music festivals. The case study on the Volkswagen emission scandal entailed the analysis of the Facebook pages of four corporate entities, the unfolding of the scandal was investigated using the Social Set Visualizer.

In the predictive case studies, it was shown that the software tool can be utilized to not only focus on the past an generate existing information, but alos predict the future outcomes on the data at hand. Thus, the Social Set Visualizer can be used for various purposes in academia and industry. For a sports brand such as Nike, the tool was applied in the context of sales forecasting. Then, the Social Set Visualizer was applied to predict concert attendance at Roskilde festival by identifying shared social media audiences through set intersections as well as to predict crowd movements for actionable insights into optimization of crowd safety operations. Lastly, a comprehensive study of the 2017 German federal election was performed using the Social Set Visualizer, which discovered that the results of this election have not directly been influenced by the level of Facebook activity for each party.

Chapter 6

Discussion

This chapter presents reflections on the utilized research methodology, the visualization of large-scale sets, the developed IT artifact, the theoretical data model, and the domain-specific query language. Furthermore, it contains a detailed discussion of implications and limitations of the research presented in this thesis.

6.1 Reflections on the Research Methodology

This section discusses the use of Action Design Research methodology in particular. Furthermore, it reflects on the lack of comparative studies between Social Set Analysis and Social Network Analysis, the data collection pipeline of the Social Set Visualizer, and the application of Social Set Analysis to non-Facebook datasets.

6.1.1 Use of Action Design Research Methodology

The work in this PhD project is grounded in **Action Design Research methodology** which relies on an iterative research process. Thus, this thesis used the initial formulation of the research problems as a starting point. Thereby, a particular focus was set on the current challenges in Big Data Analytics which were identified by the National Academy of Sciences [National Research Council *et al.* 2013]. This step was accompanied by the identification stage of the research questions concerning Visual Analytics of Big Social Data and the application of the Social Set Analysis approach to Big Social Data Analytics. Subsequently, during the building, intervention, and evaluation (BIE) stage of Action Design Research, three versions of the Social Set Visualizer IT artifact have been designed and implemented. These versions have been presented in **Publication I** [Vatrapu *et al.* 2016], **Publication II** [**Flesch** *et al.* 2015a], and **Publication III** [**Flesch** *et al.* 2016] of this dissertation. Furthermore, the novel software tools' capabilities to generate insights were evaluated through various descriptive and predictive case studies.

The IT-dominant BIE stage of Action Design Research relies on the researchers to **distribute alpha and beta versions of the IT artifact to practitioners and end-users**. In this PhD project, the alpha version was provided to fellow academics who were not part of the development of the software tool. Furthermore, the beta version was given to university students to generate insights from Big Social Data in their course projects. It has to be stated that this obsersation does not reflect the software industry, in which "end-user" demographics rarely consist of a majority of highly-educated students. Hence, a slight bias might have been introduced during the case

study evaluation of the Social Set Visualizer, as Action Design Research methodology was utilized with **academic staff and students instead of industry practitioners and real-world end users**.

6.1.2 Social Set Analysis vs. Social Network Analysis Studies

In this thesis, a brief qualitative comparison between Social Set Analysis and Social Network Analysis is presented. However, it can be argued that **no exhaustive comparative study** between Social Set Analysis and the historically dominant approach of Social Network Analysis has been conducted using the Social Set Visualizer software tool. There are several underlying reasons for the presentation of a qualitative comparative evaluation. First, a comprehensive reasoning why Social Set Analysis complements Social Network Analysis and why it is a worthwile undertaking to invest valuable research time into Social Set Analysis is given. This line of reasoning has been argued over various publications to great extent, including **Publication I** [Vatrapu *et al.* 2016] at the beginning of this dissertation. Second, the case studies based on the Social Set Analysis methodology, which were performed during this PhD project, resulted in a variety of novel research findings that needed to be documented and published. Due to this observation, publishing of novel research findings was prioritized over implementing replication studies of Social Network Analysis findings. Still, a backlog of ready-to-publish research findings that were gathered by means of Social Set Analysis has been steadily building up over the past three years.

6.1.3 Integrated Data Collection

The methodology of data collection utilized in this PhD project is deeply integrated into the Social Set Visualizer. This integration has been introduced in the third version of the tool, which is presented in **Publication IV** [**Flesch** *et al.* 2017], whereas previous versions relied on an external source of data such as SODATO, as presented in section 2.4.2. In combination with the Visual Analytics functionality, a large, continuous stretch of the Big Data Value Chain was directly implemented in the latest version of the Social Set Visualizer software tool. This results in a very streamlined acquisition and analysis of Facebook data with tangible benefits for the insight generation from Big Social Data. At the same time, this **specialization on the integrated collection of Facebook data depicts a limitation**. A data collection functionality for non-Facebook data has not been implemented in the Social Set Visualizer due to lack of API access to other leading social networks such as Instagram.

6.1.4 Social Set Analysis of non-Facebook Datasets

At the beginning of this dissertation, it was shown that this PhD project contributes a significant amount of research on large-scale Facebook datasets to the state of the art in Visual Analytics of Big Social Data. In my presented publications, the novel set-based approach to Big Social Data Analytics, Social Set Analysis, was utilized. This shows a methodological limitation of the work presented in this thesis, as all

Social Set Analysis studies have been performed on Facebook datasets. **Social Set Analysis of non-Facebook datasets is not explicitly investigated** in this PhD project, even though claims about methodological applicability to other types of Big Social Data are made.

6.2 Reflections on the Visualization of Large-scale Sets

This section reflects on the visualization of large-scale sets. First, limitations of set visualizations such as Euler and Venn diagrams are discussed. Then, the contribution of a novel type of Venn diagram used in SoSeVi 1, the "exploded" Venn diagram, and its comparison with the EulerAPE approach is examined. Furthermore, the evaluation of the three different visualizations used in the different versions of the Social Set Visualizer is discussed.

6.2.1 Limitations of Set Visualizations

The various limitations of classical set-theoretical visualizations such as Euler and Venn diagrams with regard to area-proportionality and the ability to display large-scale sets were thoroughly investigated. Furthermore, the potential of the **matrix-based UpSet approach to set visualization** is outlined through its implementation in versions two and three of the Social Set Visualizer, which were presented in **Publication III** [**Flesch** *et al.* 2016] and **Publication IV** [**Flesch** *et al.* 2017] of this dissertation. As the challenge of an area-proportional visualization of more than three sets has been resolved, its application to the set-based approach to Big Social Data Analytics enables more complex analytical studies.

6.2.2 "Exploded" Venn Diagrams and EulerAPE

The "exploded" Venn diagrams presented in **Publication II** [**Flesch** *et al.* 2015a] depict a novel solution to the problem of visualizing three sets in an area-proportional way. Compared to the EulerAPE approach [Micallef & Rodgers 2012], the novel Venn diagram type **improves the visual clarity of set intersections and digestability of the displayed data labels**. For the purpose of Visual Analytics, the "exploded" Venn diagram was shown to be a suitable replacement for EulerAPE and also the traditional Venn diagram.

6.2.3 Quantitative Ranking of Set Visualizations

No quantitative selection of the "best" set visualization is made from the three presented versions of the IT artifact, which can be seen as another limitation. Hence, the **"best" version of the Social Set Visualizer is not experimentally determined**. Even though the empirical case studies demonstrate utility of all three versions of the Social Set Visualizer for research purposes, such as the generation of research findings from Big Social Data, this thesis does not provide an exhaustive comparative

study between these versions. This lack of experimental verification is justifiable, given the already large scope of this PhD project.

6.3 Reflections on the IT Artifact

This section reflects on the IT artifact, in particular on the insight generation from Big Social Data, the feasibility of implementing an IT artifact for Big Social Data Analytics, and the impact of open-source software on development of the IT artifact. Furthermore, it discusses the utilization of Social Set Analysis by other institutions, the choice of the underlying database, and the level of interactivity in the visualizations presented for the three versions of the Social Set Visualizer software tool.

6.3.1 Generation of Insights from Big Social Data

Computational Social Science research has reached a point where social media activity is ubiquitous, yet hard to collect and to analyze in domain-specific ways. In conjunction with complex event timelines, the data at hand presents **numerous opportunities for attaining deep insights**, but many of those insights are **very difficult to uncover**. In this context, Visual Analytics presents the means of generating those insights through easily usable software tools such as the Social Set Visualizer for users with different backgrounds, both experts and novices alike.

Current state of the art approaches to Big Social Data Analytics still only scratch at the proverbial surface of potential research findings. Social Set Analysis and Social Network Analysis mostly work with metadata analysis, and are limited in that regard. Until now, research has **not found a way to perform deep analysis of actual social media content** such as political discussions, apart from vague methods like sentiment analysis. Therefore, the Social Set Visualizer and its implementation of the **set-based methodology depicts an important stepping stone** along this journey of generating more insights from Big Social Data.

6.3.2 Feasibility of Implementing an IT Artifact

The implementation of the presented Social Set Visualizer dashboard showcases that the creation of a Visual Analytics software tool, which meets the high technical, analytical and user experience requirements of present-day computing, **is viable and can be achieved by an academic researcher with limited resources** by following an iterative development approach that is grounded in Action Design Research methodology.

6.3.3 Productivity Increases through Use of Open-Source Components

Furthermore, the presented software tool leverages open-source components. This is only possible due to the high quality of the open source software projects involved, such as Linux, Ubuntu, PostgreSQL, Redis, Node.js, D3.js, React, Nginx, and many

more. The gained productivity is fueled by two factors. First, most **open-source components are free** for non-commercial use, reducing the costs incurred during development of the IT artifact. Second, due to availability of the underlying source code of open-source tools, potential software errors can easily be fixed by the researchers **without need for third-party support**.

6.3.4 Utilization of Social Set Visualizer by Other Researchers

Even though the Social Set Visualizer was evaluated in seven case studies with various researchers, these **evaluations are mainly limited to academics and students from Copenhagen Business School**. The utility of the Social Set Visualizer should be further evaluated through research projects with additional outside institutions. During the Several institutions have already expressed their interest in research collaborations based on the Social Set Visualizer software tool. However, such a collaboration was not implemented yet.

6.3.5 Choice of Databases

Additionally, it can be argued that the choice of PostgreSQL as storage backend of the Social Set Visualizer is **not a real "Big Data" solution**, as it does not represent a **distributed, parallelizable database** such as Apache Spark or Hadoop. To counter this argument, it has to be pointed out that PostgreSQL is a very mature database with many means of evaluating and fine-tuning the database performance. On top of that, the PostgreSQL maintainers are highly agile and frequently publish new features, such as JSON/BSON support, replication and performance optimizations. Furthermore, due to the decline in prices for RAM, disk space and computing, PostgreSQL can easily be configured to accomodate all datasets that are involved in this thesis. Hence, **no practical reason exists that requires the Social Set Visualizer to switch away from PostgreSQL** and use a distributed database.

6.3.6 Interactivity of Visualizations

The third version of the Social Set Visualizer, which is presented in Publication IV [Flesch *et al.* 2017] of this dissertation, contains **less interactive user interface components as the previous versions**. Due to the **substantial changes of the user interface** during implementation of the UpSetR-inspired set visualization, less interactive elements were included in the third version. The visualization of large-scale sets in a visually comparable, area-proportional way is the largest contribution of the third version of the Social Set Visualizer. Therefore, and due to time constraints, no additional interactivity was created for visualizations and user interface elements.

6.4 Reflections on the Conceptual Data Model

This section reflects on the theoretical model of Big Social Data which was developed during the PhD project, namely the Social Interaction Model. Hence, it discusses the differences to the existing Social Data Model, the potential overfitting of the Social Interaction Model to the Social Set Visualizer, and the influence of Facebook datasets on model development.

6.4.1 Differences to Existing Social Data Model

The Social Interaction Model is compared with the existing Social Data Model. Key differences between both models are presented and discussed.

Introduction of Interactions

In applied Social Set Analysis, a notion of interactions is needed. This consists of a data structure comparable to a linked, timestamped list of one initial *Action* performed by a social media *Actor* onto a target social media *Actor*, and zero or many *Reactions* to that initial action. Therefore, the Social Interaction Model strives to provide a simple, unified data structure for Social Set Analysis, inspired by Big Social Data research in real-world scenarios.

Focus on Social Interactions

The proposed model focuses on the core of socio-technical interactions between human beings over the internet. The previously established Social Data Model puts the theme of *Conversations* on the same level as *Interactions*, whereas the Social Interaction Model argues that *Conversations* are resulting from *Interactions*. Therefore, emphasis of the theoretical model should be placed on the concept of *Interactions*.

Due to the depriorization of *Social Text* and *Conversations*, it needs to be discussed whether the Social Interaction Model is going too far. By elevating the notion of *Interactions* to a core principle of the model without proving much rationale for this step, apart from a healthy intuition that it simplifies the Social Set Analysis research approach and makes it easier to implement in an IT artifact. Future research might underline that the Social Interaction Model depicts a sub-model of the Social Data Model, with no means to replace the existing model as a whole.

Definition of Spatial and Temporal Dimensions

To formalize Social Set Analysis methodology, two dimensions are needed, one dimension for time and one dimension for space. The formal definition of the existing Social Data Model can be extended in order to support these two dimensions of *Location in Space* and *Location in Time*. This set-enabled dimensionality has been included in the formal definition of the Social Interaction Model.

The introduction of temporal and spatial dimensions as foundational components streamlines analytical tasks of Big Social Data under the Social Set Analysis methodology. Furthermore, it makes the model more opinionated towards the utilization of

a certain methodology, namely Social Set Analysis. With such a bias built into the model itself, it can be argued that it loses its applicability to other methodological approaches of Computational Social Science such as Social Network Analysis. A counterargument to this observation could be that the original Social Data Model never presented an addition or contribution to Social Network Analysis methodology. Due to the same publication time as the Social Set Analysis approach, it was always intended to act as catalyst for the application of Social Set Analysis in Big Social Data Analytics.

Addition of Non-Textual "Artifact" Content Types

The existing Social Data Model lacks support for non-textual *Artifact* content types apart from *Conversations*, such as images and videos within social data. This is rectified by the proposed model through addition of *Social Images* and *Social Videos*, from which meaningful information can be extracted by utilizing state-of-the-art machine learning approaches such as deep learning.

Therefore, it is not required to conceptually attach information on *Topics*, *Keywords*, *Pronouns* and *Sentiments* to the *Conversation* domain as seen in the Social Data Model, but this information may be attached to *Artifacts* of any content type as proposed by the Social Interaction Model.

Unification of Bipartite Social Data Model

The proposed Social Interaction Model unifies the bipartite Social Data Model into one sequential concept which is based on a set-based definition of the social data and interactions. A slight refinement in the definition of *Artifacts* essentially enables the proposed model to express that meaning with respect to *Topics*, *Keywords*, *Pronouns*, and *Sentiments* can be extracted from the *Artifact* data. With *Conversations*, now called *Social Text*, depicting one specific type of *Artifact* data, we observe that there is no inherent conflict between both models, but rather that the Social Interaction Model depicts a logical extension and generalization of the Social Data Model.

Depreciation of Activities

Activities as defined in the Social Data Model are a vague concept that provides no clear mapping to the real-world datasets used for Social Set Analaysis purposes. The formal definition of *Activities* in the original publication of the Social Data Model [Mukkamala *et al.* 2013] concerns a mapping function from *Artifacts* to *Activities*, and from the presented example, an *Activity* is a "promotion" of products by a clothing retailer on Facebook, that may span over many Actors, Actions, and Artifacts. Thus, the notion of *Activity* in the Social Data Model aims to **capture the goal or intention** of a social media *Actor* that is behind their *Action* to broadcast a certain *Artifact* to the social network. It is difficult to capture the underlying goals and intentions of an *Actor* in data for research purposes.

Alternatively, when we take into account the activity theoretic notion of *Activities*, in which *Activities* depict **unconscious motivations and fundamental needs** rather

than conscious goals [Kaptelinin & Nardi 2006], it also becomes apparent that information on the unconscious intention of the *Actor* is very difficult to obtain for empirical studies.

Due to the empirical difficulties in acquiring data which exist for both presented interpretations of the term *Activities*, the concept has not been included in the Social Interaction Model.

Improved Interoperability between Data Sources

Based on the dimension of *Location in Space* which is proposed in the Social Interaction Model, it is conceptually possible to interoperate between *Artifacts* from different social media data sources. For example, after a dimensionality reduction to the temporal dimension *Location in Time*, *Artifacts* from multiple data sources such as Twitter and Facebook can be grouped and compared for Social Set Analysis purposes.

6.4.2 Overfitting of Social Interaction Model to Social Set Visualizer

The Social Interaction Model presents an incremental contribution to Social Set Analysis, as it extends upon the preexisting Social Data Model. It builds on the **learnings gathered throughout this PhD project**, and its **specifications are adapted to the special use cases of the Social Set Visualizer and the Social Set Query Language**. Therefore, one could argue that the proposed Social Interaction Model is overfitted to the Social Set Visualizer IT artifact. This argument needs to be investigated in future studies, which will show the utility of the Social Interaction Model and whether it can persist on its own without the Social Set Visualizer. However, the parallel, iterative development of both the theoretical model and the IT artifact during this PhD project is well-grounded in the state of the art. Additionally, all scientific contributions have been reviewed by academic peers.

6.4.3 Model Developed with Facebook Datasets

The formalization of the Social Interaction Model presented in this dissertation arises from **exclusive work with Big Social Data from Facebook**. Hence, it is possible that the model is too focused on this data source, and has not enough influence from other sources such as Twitter or even upcoming decentralized social networks such as Mastodon to reflect the full theoretical spectrum of Big Social Data. Even though the author is highly confident that the Social Interaction Model is applicable to other datasets, this needs to be further investigated through future research.

6.4.4 Using the Model to Express Other Forms of Online Communication and Collaboration

An application of the Social Data Model to other forms online communication and collaboration is theoretically possible, even though it stretches the definition of Big Social Data as it was utilized in this thesis.

The presented model fits to **document-based online collaboration environments** such as Wikipedia and Google Docs, in which the initial *Action* of creating a new document or article by one individual *Actor* is the reference point that spawns an textual *Artifact*. After the *Artifact* exists, other users, e.g. *Actors*, in the online system can react to the creation of the textual *Artifact* through means of a *Reaction*, which in turn spawns a modified textual Artifact that depicts the current state of the document.

Furthermore, communication within **real-time chat rooms and online discussion forums** can also be expressed using the Social Interaction Model. In an online chat room such as IRC or a discussion forum such as Twitter, many *Actors* meet. A conversation is initiated by one *Actor* interacting either with the chat room itself (e.g. as an *Actor* of type entity) or directing a conversation at another user (e.g. an *Actor* of type user). The message that is exchanged depicts a textual *Artifact*, and other *Actors* within the chat room or discussion forum are able to react to the *Action* through means of a *Reaction*.

These examples outline the theoretical applicability of the Social Interaction Model to many forms of online communication and collaboration.

6.4.5 Geospatial Set Analysis

In my most recent publication [Flesch *et al.* 2018], the concept of Social Set Analysis was applied to a large-scale GPS dataset from a multi-day music festival in Denmark. The paper surveyed a use of set-based intersection techniques to track the aggregate movement of concergoers between different stages of the festival and the camping areas. Sets are defined by geofencing a certain area and adding a time dimension, e.g. the set of all GPS devices that were at the camping site between 9pm and 10pm, and can be compared with other sets. Based on unique device IDs a set intersection can be calculated from which migration patterns in the festival area emerge. Overlaps in audience between different bands can be identified and highlighted, which enables us to answer a new set of research questions. This showcases the potential of applying the Social Set Analysis methodology to Big Data Analytics concerning **different formats and structures apart from Big Social Data** from platforms such as Facebook.

6.5 Reflections on the Domain-specific Query Language

Reflections on the Social Set Query Language, a domain-specific query language for the Social Set Visualizer which was introduced in Publication IV [Flesch *et al.* 2017], are outlined in this section.

6.5.1 Features of Social Set Query Language

The Social Set Query Language presented in this thesis exhibits a variety of features, thus potential shortcomings should be discussed. One could argue that the textual query language, as it is proposed in this dissertation, is **not powerful enough** to

cover the whole application domain of Social Set Analysis. It should be implemented as a **fully-fledged query language** such as SQL that is based on relational algebra, and is not limited to mapping a JSON-style query object towards certain database queries and set operations. Therefore, future work is needed to expand upon the ideas set forth with the presented query language.

6.5.2 Improvement of Structure and Usability

The Social Set Query Language enables a **decoupling between client- and server-side** query languages. The client side can then utilize a simplified syntax to define sets for the purpose of performing set-based Big Social Data Analytics according to the Social Set Analysis methodology. This **massive simplification** of the analytical process has **usability benefits for the end user** of the software tool. Even though such an improvement of usability has measurable overhead, the implementation of set operations in plain SQL is very verbose and requires many repeating formulations in order to achieve performance and scalability of query results. Therefore, the introduction of Social Set Query Language is a net benefit for the user.

6.5.3 Quality of Insights from Big Social Data

The Social Set Query Language improves the quality of insights from Big Social Data by quantifying findings from set-based calculations. Due to this, the contribution of the Social Set Query Language towards the creation of a focused Social Set Analysis dashboard can not be understated. It **streamlines developmental tasks** on the visualization interface, because data processing is decoupled from data visualization. Furthermore, it provides a textual formalization of the utilized research methodology which increases the **repeatability of studies** that have been performed with the Social Set Visualizer.

6.5.4 Evaluation of Social Set Query Language

The current implementation of the Social Set Query Language **does not cover all use cases** that are theoretically possible based on the presented Social Interaction Model. This is due to the fact that during the iterative Action Design Research approach of this PhD project, the Social Set Query Language was thoroughly evaluated on a technical level with regard to the data storages used. Due to the relatively recent presentation of the query language in the third version of the Social Set Visualizer, a more thorough evaluation of the Social Set Query Language should be pursued in future research.

6.5.5 Choice of PostgreSQL as Data Storage

Lastly, the Social Set Query Language requires data according to a predefined database schema that is based on the Social Interaction Model. As evaluated earlier, the **choice of storage backend** massively influences this requirement. With other

databases as storage backend such as Apache Spark, it would have been possible to be more flexible in terms of data validation and pre-processing requirements. In order to focus on the core of this PhD project, which is the Social Set Visualizer, a decision was made to operate in a **strictly defined environment with a strong database schema** based on the relational database PostgreSQL. Thereby, data conversion overhead is mitigated by implementation of a built-in social media crawler in the Social Set Visualizer, so that in many cases no external data files need to be loaded and converted into the internal database schema.

6.6 Summary

This chapter discussed the findings and respective outputs presented in this dissertation. First, reflections on research methodology were presented, with special focus on the suitability of Action Design Research methodology. This methodology particularly fits to this PhD project due to its iterative approach involving alpha and beta versions used by academic stakeholders. Furthermore, the theoretical and practical comparison between Social Set Analysis and Social Network Analysis resulted in various research findings from use of the novel Social Set Analysis methodology during the course of this thesis, although no exhaustive comparative study between both has been presented. Moreover, the integration of a data collection pipeline in SoSeVi 3 depicts a step towards implementation of the Big Data Value Chain, but also limits the tool to datasets collected from Facebook. Thus, conducting of Social Set Analysis studies with non-Facebook datasets is not explicitly investigated in this PhD project. Reflections on the visualization of sets are provided with special regard to the challenge of area-proportional set visualization using traditional Euler and Venn diagrams. The introduction of "exploded" Venn diagrams in SoSeVi 1 presents a novel alternative to EulerAPE, which has limitations in terms of readability and visual consistency. Further, various limitations of explicitly ranking the presented visualization types, due to lack of agreed-on objective criteria, were discussed. Additionally, the IT artifact was discussed in detail, with particular focus on insight generation, implementation feasibility, database choice, and the interactivity of user interfaces. Lastly, extensive reflections were made on the presented Social Interaction Model and the domain-specific textual query language for Social Set Analysis.

Chapter 7

Conclusions and Future Work

This chapter summarizes the findings of this PhD project and concludes the dissertation in view of the presented research questions. First, practical and theoretical contributions of this PhD project are highlighted. Then, a conclusion on the main research questions is drawn and put forward. The chapter closes with an outline of future work which should be considered to further advance the state of the art in research on Visual Analytics of Big Social Data.

7.1 Contributions

This dissertation contributed to Big Data Analytics and Computational Social Sciences by providing novel solutions to the two key challenges of ***"working with different data formats and structures"*** and ***"developing methods for visualizing massive data"*** identified by the National Academy of Sciences's report on Massive Data Analysis [National Research Council *et al.* 2013].

It contributed the Social Interaction Model, a conceptual model of Big Social Data that streamlines and set-theoretically extends the previously used Social Data Model [Mukkamala *et al.* 2013, Vatrapu *et al.* 2016]. The Social Interaction Model **combines different data formats and structures in a unified theoretical model** for Big Social Data, which benefits data exploitation and use of Social Set Analysis methodology. The utility of the conceptual model and the flexibility of the analytical processes were demonstrated with various large-scale Facebook and GPS datasets.

Furthermore, this PhD project contributed the Social Set Visualizer IT artifact, which depicts the **first tool to use set-based visualizations** for Visual Analytics of Big Social Data and also the **first tool to utilize Social Set Analysis** methodology for insight generation from Big Social Data. Following the Action Design Research methodology, this dissertation provided **extensive documentation on design and development** of three iterative versions of the Social Set Visualizer. In addition, it contributed an **evaluation consisting of seven case studies** which utilized the Social Set Visualizer in descriptive and predictive analytics of real-world problems in several industries.

The publications presented in this thesis **significantly increased the number of data points** used in state-of-the-art research in Visual Analytics of Big Social Data. This increase was by 100x in relation to the mean and by 10,000x in relation to the median size of prior research on Facebook datasets. Thereby, dataset sizes of state-of-the-art research focusing on Facebook datasets have been elevated to the same level as research using Twitter datasets. **Closing of this sizeable research**

gap was enabled by a novel set-based approach to data exploitation in the Social Set Visualizer. Data exploitation was improved through application of Social Set Analysis, which **resolved important theoretical and methodological limitations** in Big Social Data Analytics.

This PhD project contributed several solutions the key challenge of *"developing methods for visualizing massive data"*. Overall, **three innovative visualizations** were contributed to the field of Big Social Data Analytics, namely "Exploded" Venn diagrams, UpSet- and UpSetR-styled visualizations. For each of these visualizations, a **dynamic, interactive, and browser-based implementation** in the Javascript programming language was provided.

A **novel area-proportional three-set visualization**, the "Exploded" Venn diagram, was designed and developed in this thesis. It is very suitable for use in interactive Visual Analytics due to its visual consistency and the clarity of its labels. Therefore, the "Exploded" Venn diagram depicts a **distinct contribution to the field of Visual Analytics** on its own.

Furthermore, this dissertation is the first to utilize **UpSet- and UpSetR-style visualizations** from the field of bioinformatics for the generation of insights from Big Social Data. Likewise, this dissertation contributed to the advancement of the UpSetR approach to set visualization by **introducing logarithmic scales, shading, and color coding**, thereby signifying the number of set intersections displayed in the bar chart and individual sets in the combination matrix.

Lastly, this thesis contributed the Social Set Query Language, which depicts the **first textual query language for Social Set Analysis** of Big Social Data. It enables formalization and documentation of set-based research studies, thereby **increasing reproducibility and quality of insights** in Big Social Data Analytics. The utilization of the query language within the Social Set Visualizer dashboard and its impact on simplification of client-to-server communication was showcased. Furthermore, an **extensive evaluation of different databases** with suitability for implementation of the Social Set Query Language was given and the resulting decision for a relational database was presented.

Contributions to Social Set Analysis

Three core contributions have been presented with particular relevance for Social Set Analysis. These contributions add both to the theoretical and practical foundations of Social Set Analysis, as illustrated in Figure 7.1.

First, the **Social Interaction Model** contributes a theoretical extension to the two existing versions of the Social Data Model [Mukkamala *et al.* 2013,Vatrapu *et al.* 2016]. Thereby, it advances and replaces the Social Data Model as the foundational theoretical data model for Social Set Analysis methodology.

Second, the **Social Set Visualizer** software tool constitutes the major practical contribution of this PhD project. It extends the Social Graph Analytics Tool (SOGATO) [Hussain & Vatrapu 2011] and the Social Data Analytics Tool (SODATO) [Hussain & Vatrapu 2014b], which present the two previously developed IT artifacts for Big

Social Data Analytics from members of our research group. In contrast to the two previous tools, the Social Set Visualizer depicts the first Visual Analytics tool to implement the Social Set Analysis approach. Furthermore, it provides capabilities of insight generation for Big Social Data Analytics through UpSet- and UpSetR-style visualizations of large-scale sets and set intersections.

Third, the **Social Set Query Language** successfully bridges the theoretical and practical realms of this PhD project by linking the Social Interaction Model and the Social Set Visualizer IT artifact. As it depicts a simple, domain-specific textual query language for Social Set Analysis, it allows researchers to formalize their set-based studies through a textual definition of sets. Thereby, it increases reproducibility of studies and auditability of findings.

The **field of Social Set Analysis is significantly advanced** by the three core contributions of this dissertation. Furthermore, its theoretical and practical pillars are merged in a unified analytical platform. The Social Set Visualizer software tool is the first tool to directly incorporate the theoretical data model of Big Social Data in the insight generation process. Through use of the Social Set Query Language, analytical studies are formulated as set-based queries and visualized using the Visual Analytics tool.

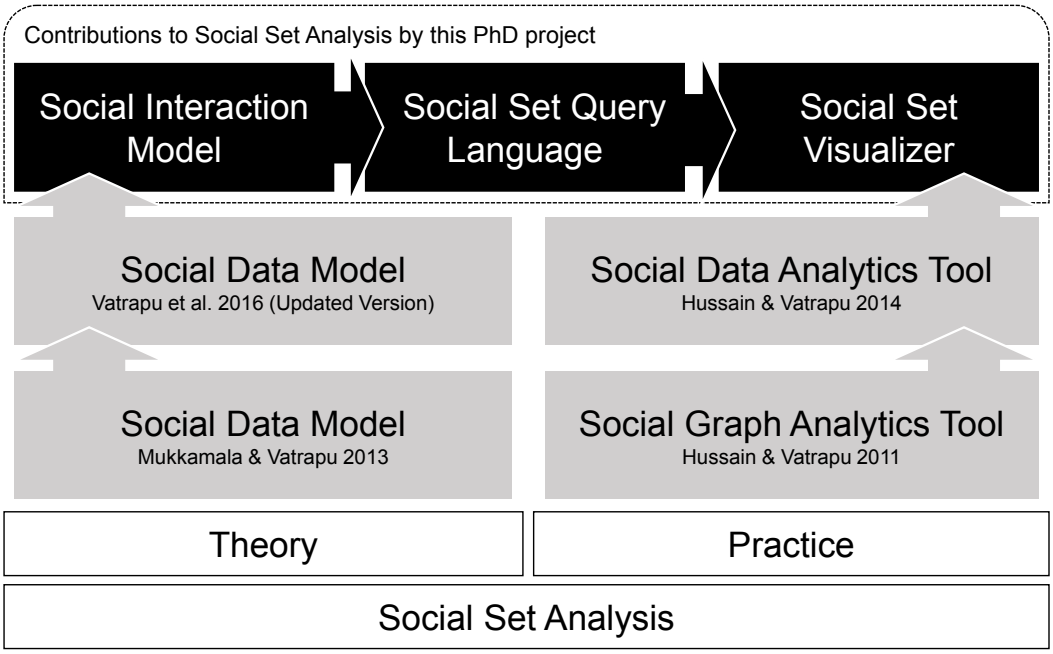


Figure 7.1: Illustrative overview of this dissertation's contributions to theory and practice of Social Set Analysis

7.2 Conclusions

In conclusion of this dissertation, the two initial research questions are revisited and answers to them are put forward.

> RQ1: *How and in what way can the novel Social Set Analysis approach to Big Social Data Analytics be modeled into an interactive Visual Analytics software tool that can be utilized for generating meaningful insights from Big Social Data?*

The first research question of this thesis concerns the feasibility of creating an interactive Visual Analytics software tool which incorporates the Social Set Analysis approach to Big Data Analytics, with special focus on the generation of meaningful insights from Big Social Data.

In this PhD project, the design, development, and evaluation of three different versions of the Social Set Visualizer was presented. Hence, the general feasibility of a Visual Analytics software tool that is grounded in Social Set Analysis methodology has been successfully demonstrated. Furthermore, seven different case studies were provided in which the software tool was utilized to generate insights from a diverse set of Facebook datasets. The meaningfulness of the insights generated using the Social Set Visualizer is underlined by the fact that five out of the seven presented case studies have been peer-reviewed and published at relevant conferences.

> RQ2: *What are software design requirements for a Visual Analytics software tool that interactively visualizes large-scale sets and set intersections with multiple users and large amounts of data?*

The second research question of this thesis concerns the software design requirements of the Social Set Visualizer which enable its interactive visualization of large-scale sets and set intersections for multiple users and large amounts of data.

The research performed in this PhD project has provided evidence that the Social Set Visualizer facilitates the analysis of significantly larger Facebook datasets than the state of the art in Visual Analytics of Big Social Data. Moreover, the three different versions of the Social Set Visualizer incorporate three innovative approaches to set-based visualizations, namely the creation of the novel "Exploded" Venn diagram alongside the application and refinement of the UpSet- and the UpSetR-style set visualizations from the field of bioinformatics for use in Big Social Data Analytics. For a target audience that consists of researchers and practitioners, usability was identified as the core design objective for the Social Set Visualizer, consisting of five components efficiency, learnability, memorability, user satisfaction, and error handling. The multi-user requirement was resolved by creating a web-based Visual Analytics dashboard that can be accessed simultaneously using a web browser without the need to install additional software. Hence, visualizations of large-scale sets and set intersections are dynamically displayed in the client-side component of

the Social Set Visualizer within the user's web browser, while the actual set computations are performed in the server-side component. After rigorous evaluation of potential storage backends suitable for implementation of the Social Set Analysis methodology, a relational database, PostgreSQL, was chosen due to its mature optimization features, availability of the SQL query language, and incorporation of the theoretical model of Big Social Data, the Social Interaction Model, into a well-defined database schema. Furthermore, the positive effect of caching on overall performance of the Social Set Visualizer was demonstrated through use of Redis as ephemeral in-memory database.

7.3 Future Work

A variety of future work approaches have been elaborated in the course of this PhD project. In conclusion of this dissertation, four major streams of future work have been identified as key focus areas resulting from the findings of this thesis.

First, the utility of Social Set Analysis needs to be **empirically demonstrated with non-Facebook datasets**. For a long time, our research group has mainly relied on the good access to Big Social Data from Facebook, from which a multitude of research findings could be generated and published in journals and conferences. In order to provide further empirical proof of the applicability of Social Set Analysis to interesting research problems and diverse sets of social media data, studies based on other data sources need to be created and published.

Second, the Social Set Visualizer is still basically a 2D Visual Analytics dashboard. In face of emerging technologies such as **virtual reality, mixed reality and augmented reality** in products such as the Microsoft Hololens, it should be explored in how far UpSetR-style large-scale set visualizations can be implemented and visualized in 3D space. During this PhD project, several mixed reality prototypes have been built, but no conclusive results were found.

Third, the topic of **geospatial set analysis** is still largely unexplored. A set-based approach to geospatial analytics has been demonstrated in a recent publication [**Flesch** *et al.* 2018], however further depth needs to be developed and expanded. Like the Social Set Visualizer, a specialized IT artifact for performing geospatial set analysis could be designed, developed and evaluated.

Fourth, the design and development of a **custom database tailormade for Social Set Analysis** with a direct implementation of the Social Set Query Language without SQL as an intermediary should be further researched. This could allow a way to gain additional operational performance for the Social Set Visualizer tool. Future work on this issue might exhibit a stronger focus on software engineering and database design.

Bibliography

[Abbasi & Chen 2007] Ahmed Abbasi and Hsinchun Chen. *Categorization and Analysis of Text in Computer Mediated Communication Archives Using Visualization*. In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '07, pages 11–18, New York, NY, USA, 2007. ACM. (Cited on page 26.)

[Abbasi & Chen 2008] Ahmed Abbasi and Hsinchun Chen. *CyberGate: A Design Framework and System for Text Analysis of Computer-Mediated Communication*. MIS Quarterly, vol. 32, no. 4, pages 811–837, 2008. (Cited on page 26.)

[Abbasi *et al.* 2013] Ahmed Abbasi, Tianjun Fu, Daniel Zeng and Donald Adjeroh. *Crawling credible online medical sentiments for social intelligence*. In Social Computing (SocialCom), 2013 International Conference on, pages 254–263. IEEE, 2013. (Cited on page 26.)

[Abras *et al.* 2004] Chadia Abras, Diane Maloney-Krichmar and Jenny Preece. *User-centered design*. Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications, vol. 37, no. 4, pages 445–56, 2004. (Cited on page 35.)

[Acharya & Park 2016] Srijana Acharya and Han Woo Park. *Open data in Nepal: a webometric network analysis*. Quality & Quantity, pages 1–17, 2016. (Cited on page 190.)

[Albert & Tullis 2013] William Albert and Thomas Tullis. Measuring the user experience: collecting, analyzing, and presenting usability metrics. Newnes, 2013. (Cited on page 53.)

[Alsallakh *et al.* 2016] Bilal Alsallakh, Luana Micallef, Wolfgang Aigner, Helwig Hauser, Silvia Miksch and Peter Rodgers. *The State-of-the-Art of Set Visualization*. In Computer Graphics Forum, volume 35, pages 234–260. Wiley Online Library, 2016. (Cited on page 46.)

[Archambault & Hurley 2014] Daniel Archambault and Neil Hurley. *Visualization of trends in subscriber attributes of communities on mobile telecommunications networks*. Social Network Analysis and Mining, vol. 4, no. 1, pages 1–17, 2014. (Cited on page 187.)

[Armbrust *et al.* 2015] Michael Armbrust, Reynold S. Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K. Bradley, Xiangrui Meng, Tomer Kaftan, Michael J. Franklin, Ali Ghodsi and Matei Zaharia. *Spark SQL: Relational Data Processing in Spark*. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, SIGMOD '15, pages 1383–1394, New York, NY, USA, 2015. ACM. (Cited on page 39.)

[Atkins *et al.* 1999] D. L. Atkins, T. Ball, G. Bruns and K. Cox. *Mawl: a domain-specific language for form-based services*. IEEE Transactions on Software Engineering, vol. 25, no. 3, pages 334–346, May 1999. (Cited on page 39.)

[Bello-Orgaz *et al.* 2016] Gema Bello-Orgaz, Jason J. Jung and David Camacho. *Social big data: Recent achievements and new challenges*. Information Fusion, vol. 28, pages 45–59, March 2016. (Cited on page 2.)

[ben Khalifa *et al.* 2016] Mohamed ben Khalifa, Rebeca P Díaz Redondo, Ana Fernández Vilas and Sandra Servia Rodríguez. *Identifying urban crowds using geo-located Social media data: a Twitter experiment in New York City*. Journal of Intelligent Information Systems, pages 1–22, 2016. (Cited on page 187.)

[Benjamin *et al.* 2014] Victor Benjamin, Wingyan Chung, Ahmed Abbasi, Joshua Chuang, Catherine A Larson and Hsinchun Chen. *Evaluating text visualization for authorship analysis*. Security Informatics, vol. 3, no. 1, page 1, 2014. (Cited on page 187.)

[Berkowitz & Gibbs 1979] Marvin Berkowitz and J. C. Gibbs. *A Preliminary Manual for Coding Transactive Features of Dyadic Discussion*. Ohio State University, vol. Fall, 01 1979. (Cited on page 25.)

[Binder 1998] John Binder. *The event study methodology since 1969*. Review of quantitative Finance and Accounting, vol. 11, no. 2, pages 111–137, 1998. (Cited on page 21.)

[Boiy & Moens 2009] Erik Boiy and Marie-Francine Moens. *A machine learning approach to sentiment analysis in multilingual Web texts*. Information retrieval, vol. 12, no. 5, pages 526–558, 2009. (Cited on page 26.)

[Boldt *et al.* 2016] Linda Camilla Boldt, Vinothan Vinayagamoorthy, Florian Winder, Melanie Schnittger, Mats Ekran, Raghava Rao Mukkamala, Niels Buus Lassen, **Benjamin Flesch**, Abid Hussain and Ravi Vatrapu. *Forecasting Nike's sales using Facebook data*. In Big Data (Big Data), 2016 IEEE International Conference on, pages 2447–2456. IEEE, 2016. (Cited on pages 15, 30 and 74.)

[Borgatti *et al.* 2009] Stephen P. Borgatti, Ajay Mehra, Daniel J. Brass and Giuseppe Labianca. *Network Analysis in the Social Sciences*. Science, vol. 323, no. 5916, pages 892–895, February 2009. (Cited on page 7.)

[Bostock 2012] Michael Bostock. *D3. js*. Data Driven Documents, 2012. (Cited on pages 59 and 60.)

[Boyd & Ellison 2007] Danah Boyd and Nicole Ellison. *Social Network Sites: Definition, History, and Scholarship*. Journal of Computer-Mediated Communication, vol. 13, no. 1, 2007. (Cited on page 7.)

[Bromiley *et al.* 1988] Philip Bromiley, Michele Govekar and Alfred Marcus. *On using event-study methodology in strategic management research*. Technovation, vol. 8, no. 1, pages 25–42, 1988. (Cited on page 21.)

[Carley *et al.* 2014] Kathleen M. Carley, Jürgen Pfeffer, Fred Morstatter and Huan Liu. *Embassies burning: toward a near-real-time assessment of social media using geo-temporal dynamic network analytics*. Social Network Analysis and Mining, vol. 4, no. 1, page 195, August 2014. (Cited on page 183.)

[Chae *et al.* 2014] Junghoon Chae, Dennis Thom, Yun Jang, SungYe Kim, Thomas Ertl and David S. Ebert. *Public behavior response analysis in disaster events utilizing visual analytics of microblog data*. Computers & Graphics, vol. 38, pages 51 – 60, 2014. (Cited on page 182.)

[Chae 2015] Bongsug (Kevin) Chae. *Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research*. International Journal of Production Economics, vol. 165, pages 247 – 259, 2015. (Cited on page 183.)

[Chapman *et al.* 2014] Peter Chapman, Gem Stapleton, Peter Rodgers, Luana Micallef and Andrew Blake. *Visualizing sets: an empirical comparison of diagram types*. In International Conference on Theory and Application of Diagrams, pages 146–160. Springer, 2014. (Cited on page 46.)

[Chen & Boutros 2011] Hanbo Chen and Paul C Boutros. *VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R*. BMC bioinformatics, vol. 12, no. 1, page 35, 2011. (Cited on page 43.)

[Cheng & Edwards 2015] Mingming Cheng and Deborah Edwards. *Social media in tourism: a visual analytic approach*. Current Issues in Tourism, vol. 18, no. 11, pages 1080–1087, 2015. (Cited on page 184.)

[Chow & Ruskey 2003] Stirling Chow and Frank Ruskey. *Drawing area-proportional Venn and Euler diagrams*. In International Symposium on Graph Drawing, pages 466–477. Springer, 2003. (Cited on page 43.)

[Chua *et al.* 2015] Alvin Chua, Ernesto Marcheggiani, Loris Servillo and Andrew Vande Moere. Flowsampler: Visual analysis of urban flows in geolocated social media data, pages 5–17. Springer International Publishing, Cham, 2015. (Cited on page 185.)

[Cleveland 2001] William S. Cleveland. *Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics*. International Statistical Review, vol. 69, no. 1, pages 21–26, 2001. (Cited on page 8.)

[Collins *et al.* 2004] Allan Collins, Diana Joseph and Katerine Bielaczyc. *Design research: Theoretical and methodological issues*. The Journal of the learning sciences, vol. 13, no. 1, pages 15–42, 2004. (Cited on page 17.)

[Conway *et al.* 2017] Jake R Conway, Nils Gehlenborg and Alexander Lex. *UpSetR: an R package for the visualization of intersecting sets and their properties*. Bioinformatics, vol. 33, no. 18, pages 2938–2940, 06 2017. (Cited on pages xvii, 49 and 50.)

[Crenshaw 1990] Kimberle Crenshaw. *Mapping the margins: Intersectionality, identity politics, and violence against women of color*. Stan. L. Rev., vol. 43, page 1241, 1990. (Cited on page 9.)

[Cvijikj & Michahelles 2013] Irena Pletikosa Cvijikj and Florian Michahelles. *Online engagement factors on Facebook brand pages*. Social Network Analysis and Mining, vol. 3, no. 4, pages 843–861, 2013. (Cited on page 189.)

[Dos Santos Jr *et al.* 2016] Raimundo F Dos Santos Jr, Arnold Boedihardjo, Sumit Shah, Feng Chen, Chang-Tien Lu and Naren Ramakrishnan. *The big data of violent events: algorithms for association analysis using spatio-temporal storytelling*. GeoInformatica, pages 1–43, 2016. (Cited on page 191.)

[D'hont *et al.* 2012] Angélique D'hont, France Denoeud, Jean-Marc Aury, Franc-Christophe Baurens, Françoise Carreel, Olivier Garsmeur, Benjamin Noel, Stéphanie Bocs, Gaëtan Droc, Mathieu Rouard*et al. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants*. Nature, vol. 488, no. 7410, page 213, 2012. (Cited on pages xvii and 44.)

[Emirbayer 1997] Mustafa Emirbayer. *Manifesto for a relational sociology*. The American Journal of Sociology, vol. 103(2), pages 281–317, 1997. (Cited on page 7.)

[Etezadi-Amoli & Farhoomand 1996] Jamshid Etezadi-Amoli and Ali F Farhoomand. *A structural model of end user computing satisfaction and user performance*. Information & management, vol. 30, no. 2, pages 65–73, 1996. (Cited on page 53.)

[Ferrara 2012] Emilio Ferrara. *A large-scale community structure analysis in Facebook*. EPJ Data Science, vol. 1, no. 1, page 1, 2012. (Cited on page 188.)

[Fisher *et al.* 2012] Danyel Fisher, Rob DeLine, Mary Czerwinski and Steven Drucker. *Interactions with big data analytics*. interactions, vol. 19, no. 3, pages 50–59, 2012. (Cited on page 5.)

[**Flesch** & Vatrapu 2016] **Benjamin Flesch** and Ravi Vatrapu. *Social Set Visualizer (SoSeVi) II: Interactive Computational Set Analysis of Big Social Data*. In Enterprise Distributed Object Computing Workshop (EDOCW), 2016 IEEE 20th International, pages 1–4. IEEE, 2016. (Cited on page 15.)

[**Flesch** *et al.* 2015a] **Benjamin Flesch**, Abid Hussain and Ravi Vatrapu. *Social Set Visualizer: Demonstration of Methodology and Software*. In 2015 IEEE 19th International Enterprise Distributed Object Computing Workshop, pages 148–151, Sept 2015. (Cited on pages 13, 35, 36, 45, 52, 53, 57, 60, 83, 85 and 147.)

[**Flesch** *et al.* 2015b] **Benjamin Flesch**, Ravi Vatrapu, Raghava Rao Mukkamala and Abid Hussain. *Social Set Visualizer: A Set Theoretical Approach to Big Social Data Analytics of Real-world Events*. In Big Data (Big Data), 2015 IEEE International Conference on, pages 2418–2427. IEEE, 2015. (Cited on pages xviii, 14, 30, 67 and 68.)

[**Flesch** *et al.* 2016] **Benjamin Flesch**, Raghava Rao Mukkamala, Abid Hussain and Ravi Vatrapu. *Social Set Visualizer (SoSeVi) II: Interactive Social Set Analysis of Big Data.* In SetVR@ Diagrams, pages 19–28, 2016. (Cited on pages xviii, 13, 36, 37, 48, 52, 53, 58, 61, 71, 72, 73, 83, 85 and 153.)

[**Flesch** *et al.* 2017] **Benjamin Flesch**, Ravi Vatrapu and Raghava Rao Mukkamala. *A Big Social Media Data Study of the 2017 German Federal Election Based on Social Set Analysis of Political Party Facebook Pages with SoSeVi*. In Big Data (Big Data), 2017 IEEE International Conference on, pages 2720–2729. IEEE, 2017. (Cited on pages 14, 29, 30, 37, 38, 50, 52, 53, 65, 78, 79, 80, 84, 85, 87, 91 and 165.)

[**Flesch** *et al.* 2018] **Benjamin Flesch**, Ravi Vatrapu, Raghava Rao Mukkamala and René Madsen. *Real-time Geospatial Visualization of Crowd Trajectory at Roskilde Festival 2018*. In ICIS 2018 Special Interest Group on Geographic Information Systems (SIGGIS) Pre-Conference Workshop Proceedings. 1., SIGGIS '18. ACM, 2018. (Cited on pages 15, 30, 91 and 99.)

[**Flesch** 2018] **Benjamin Flesch**. *Social Interaction Model*. In Big Data (Big Data), 2018 IEEE International Conference on. IEEE, 2018. (Cited on pages 14, 22, 24 and 177.)

[Gantz & Reinsel 2011] John Gantz and David Reinsel. *Extracting value from chaos*. IDC iview, vol. 1142, no. 2011, pages 1–12, 2011. (Cited on page 1.)

[Giatsoglou *et al.* 2016] Maria Giatsoglou, Despoina Chatzakou, Vasiliki Gkatziaki, Athena Vakali and Leonidas Anthopoulos. *CityPulse: A Platform Prototype for Smart City Social Data Mining*. Journal of the Knowledge Economy, vol. 7, no. 2, pages 344–372, 2016. (Cited on page 182.)

[Goebel & Gruenwald 1999] Michael Goebel and Le Gruenwald. *A Survey of Data Mining and Knowledge Discovery Software Tools*. SIGKDD Explor. Newsl., vol. 1, no. 1, pages 20–33, June 1999. (Cited on page 39.)

[Gottfried 2015] Björn Gottfried. *A comparative study of linear and region based diagrams*. Journal of Spatial Information Science, vol. 2015, no. 10, pages 3–20, 2015. (Cited on page 46.)

[Gratzl *et al.* 2013] Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister and Marc Streit. *Lineup: Visual analysis of multi-attribute rankings*. IEEE transactions on visualization and computer graphics, vol. 19, no. 12, pages 2277–2286, 2013. (Cited on page 47.)

[Groenli *et al.* 2018] Tor-Morten Groenli, **Benjamin Flesch**, Raghava Rao Mukkamala and Ravi Vatrapu. *Internet of Things Big Data Analytics: The Case of Noise Level Measurements at the Roskilde Music Festival*. In Big Data (Big Data), 2018 IEEE International Conference on. IEEE, 2018. (Cited on page 15.)

[Gross & Yellen 2005] Jonathan L Gross and Jay Yellen. Graph theory and its applications. CRC press, 2005. (Cited on page 7.)

[Guyot 2012] Paul Guyot. *What is the average length (in characters) of status updates on Facebook?*, December 2012. (Cited on page 6.)

[Hartmann *et al.* 2008] Jan Hartmann, Alistair Sutcliffe and Antonella De Angeli. *Towards a theory of user judgment of aesthetics and user interface quality*. ACM Transactions on Computer-Human Interaction (TOCHI), vol. 15, no. 4, page 15, 2008. (Cited on page 34.)

[Hashem *et al.* 2015] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani and Samee Ullah Khan. *The rise of "big data" on cloud computing: Review and open research issues*. Information systems, vol. 47, pages 98–115, 2015. (Cited on pages 1 and 2.)

[Heer & Agrawala 2007] Jeffrey Heer and Maneesh Agrawala. *Design Considerations for Collaborative Visual Analytics*. In IEEE Visual Analytics Science & Technology (VAST), pages 171–178, 2007. (Cited on page 182.)

[Henfridsson 2011] Ola Henfridsson. *Action Design Research*. Viktoria Institue, 2011. (Cited on page 19.)

[Hennig *et al.* 2016] Anna Hennig, Anne-Sofie Åmodt, Henrik Hernes, Helene Nygårdsmoen, Peter Arenfeldt Larsen, Raghava Rao Mukkamala, **Benjamin Flesch**, Abid Hussain and Ravi Vatrapu. *Big Social Data Analytics of Changes in Consumer Behaviour and Opinion of a TV Broadcaster*. In Big Data (Big Data), 2016 IEEE International Conference on, pages 3839–3848. IEEE, 2016. (Cited on pages xviii, 15, 30, 70 and 71.)

[Hevner 2007] Alan R Hevner. *A three cycle view of design science research*. Scandinavian journal of information systems, vol. 19, no. 2, page 4, 2007. (Cited on page 17.)

[Hussain & Vatrapu 2011] Abid Hussain and Ravi Vatrapu. *SOGATO: A Social Graph Analytics Tool*. 2011. (Cited on pages xvii, 28 and 96.)

[Hussain & Vatrapu 2014a] A. Hussain and R. Vatrapu. *Social Data Analytics Tool: Design, Development, and Demonstrative Case Studies*. In Enterprise Distributed Object Computing Conference Workshops and Demonstrations (EDOCW), 2014 IEEE 18th International, pages 414–417, Sept 2014. (Cited on pages xvii, 23, 29 and 74.)

[Hussain & Vatrapu 2014b] Abid Hussain and Ravi Vatrapu. *Social data analytics tool (sodato)*. In International Conference on Design Science Research in Information Systems, pages 368–372. Springer, 2014. (Cited on pages 4, 7, 29, 30, 69 and 96.)

[Hussain *et al.* 2014] Abid Hussain, Ravi Vatrapu, Daniel Hardt and Zeshan Jaffari. *Social Data Analytics Tool: A Demonstrative Case Study of Methodology and Software*. In Analysing Social Media Data and Web Networks. Palgrave Macmillan, 2014. (Cited on pages 10 and 29.)

[Isaksen & Bertacco 2006] Beth Isaksen and Valeria Bertacco. *Verification through the principle of least astonishment*. In Proceedings of the 2006 IEEE/ACM international conference on Computer-aided design, pages 860–867. ACM, 2006. (Cited on page 35.)

[Issa & Isaias 2015] Tomayess Issa and Pedro Isaias. *Usability and Human Computer Interaction (HCI)*. In Sustainable Design, pages 19–36. Springer, 2015. (Cited on page 34.)

[James 1987] Geoffrey James. The tao of programming. InfoBooks, 1987. (Cited on page 35.)

[Jeffrey *et al.* 2010] Heer Jeffrey, Bostock Michael and Ogievetsky VADIM. *A Tour through the Visualization Zoo*. Communications of the ACM, vol. 53, no. 6, pages 56–67, 2010. (Cited on page 5.)

[Jha *et al.* 2016] Ayan Jha, Leesa Lin and Elena Savoia. *The use of social media by state health departments in the US: analyzing health communication through Facebook*. Journal of community health, vol. 41, no. 1, pages 174–179, 2016. (Cited on page 188.)

[Kaptelinin & Nardi 2006] Victor Kaptelinin and Bonnie A Nardi. Acting with technology: Activity theory and interaction design. MIT press, 2006. (Cited on page 90.)

[Keim *et al.* 2008] Daniel A Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas and Hartmut Ziegler. *Visual analytics: Scope and challenges*. In Visual data mining, pages 76–90. Springer, 2008. (Cited on page 5.)

[Kim *et al.* 2016a] Wook-Hee Kim, Jinwoong Kim, Woongki Baek, Beomseok Nam and Youjip Won. *NVWAL: Exploiting NVRAM in Write-Ahead Logging*. SIGOPS Oper. Syst. Rev., vol. 50, no. 2, pages 385–398, March 2016. (Cited on page 182.)

[Kim *et al.* 2016b] Yongsung Kim, Eenjun Hwang and Seungmin Rho. *Twitter news-in-education platform for social, collaborative, and flipped learning*. The Journal of Supercomputing, pages 1–19, 2016. (Cited on page 189.)

[Kucher *et al.* 2015] Kostiantyn Kucher, Teri Schamp-Bjerede, Andreas Kerren, Carita Paradis and Magnus Sahlgren. *Visual analysis of online social media to open*

up the investigation of stance phenomena. Information Visualization, page 1473871615575079, 2015. (Cited on page 184.)

[Lazer *et al.* 2009] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy and Marshall Van Alstyne. *Computational Social Science*. Science, vol. 323, no. 5915, pages 721–723, February 2009. (Cited on page 8.)

[Lee *et al.* 2016] Kuo-Chan Lee, Chih-Hung Hsieh, Li-Jia Wei, Ching-Hao Mao, Jyun-Han Dai and Yu-Ting Kuang. *Sec-Buzzer: cyber security emerging topic mining with open threat intelligence retrieval and timeline event annotation*. Soft Computing, pages 1–14, 2016. (Cited on page 190.)

[Lex *et al.* 2014] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot and Hanspeter Pfister. *UpSet: visualization of intersecting sets*. IEEE transactions on visualization and computer graphics, vol. 20, no. 12, pages 1983–1992, 2014. Live Demo: http://vcg.github.io/upset. (Cited on pages xvii, 13, 47, 49, 50 and 73.)

[Li *et al.* 2016] Chenhui Li, George Baciu and Yunzhe Wang. *Module-based visualization of large-scale graph network data*. Journal of Visualization, pages 1–11, 2016. (Cited on page 189.)

[Liu *et al.* 2014] Shixia Liu, Weiwei Cui, Yingcai Wu and Mengchen Liu. *A survey on information visualization: recent advances and challenges*. The Visual Computer, vol. 30, no. 12, pages 1373–1393, 2014. (Cited on page 187.)

[Liu *et al.* 2016] Zhen Hua Liu, Beda Hammerschmidt, Doug McMahon, Ying Liu and Hui Joe Chang. *Closing the Functional and Performance Gap Between SQL and NoSQL*. In Proceedings of the 2016 International Conference on Management of Data, SIGMOD '16, pages 227–238, New York, NY, USA, 2016. ACM. (Cited on page 39.)

[Loukides 2012] Mike Loukides. What is data science? O'Reilly Media, 2012. (Cited on page 8.)

[Ma *et al.* 2016] Cui-Xia Ma, Yang Guo and Hong-An Wang. *VideoMap: An interactive and scalable visualization for exploring video content*. Computational Visual Media, vol. 2, no. 3, pages 291–304, 2016. (Cited on page 184.)

[MacKinlay 1997] A Craig MacKinlay. *Event studies in economics and finance*. Journal of economic literature, pages 13–39, 1997. (Cited on pages 21 and 22.)

[MacQueen 1967] Gailand Williard MacQueen. *The Logic Diagram*. PhD thesis, 1967. (Cited on page 42.)

[Magdy *et al.* 2014] Amr Magdy, Louai Alarabi, Saif Al-Harthi, Mashaal Musleh, Thanaa M. Ghanem, Sohaib Ghani and Mohamed F. Mokbel. *Taghreed: A System for Querying, Analyzing, and Visualizing Geotagged Microblogs*. In Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '14, pages 163–172, New York, NY, USA, 2014. ACM. (Cited on page 185.)

[Marcus *et al.* 2011] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden and Robert C. Miller. *Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11, pages 227–236, New York, NY, USA, 2011. ACM. (Cited on page 183.)

[McWilliams & Siegel 1997] Abagail McWilliams and Donald Siegel. *Event studies in management research: Theoretical and empirical issues*. Academy of management journal, vol. 40, no. 3, pages 626–657, 1997. (Cited on page 21.)

[Micallef & Rodgers 2012] Luana Micallef and Peter Rodgers. *Poster: Drawing area-proportional venn-3 diagrams using ellipses*. 2012. (Cited on pages xvii, 45 and 85.)

[Miller & Mork 2013] H Gilbert Miller and Peter Mork. *From data to decisions: a value chain for big data*. IT Professional, vol. 15, no. 1, pages 57–59, Jan 2013. (Cited on pages xvii, 3 and 50.)

[Mizruchi 1994] Mark S Mizruchi. *Social network analysis: Recent achievements and current controversies*. Acta sociologica, vol. 37, no. 4, pages 329–343, 1994. (Cited on page 8.)

[Muelder *et al.* 2014] Chris Muelder, Liang Gou, Kwan-Liu Ma and Michelle X Zhou. *Multivariate Social Network Visual Analytics*. In Multivariate Network Visualization, pages 37–59. Springer, 2014. (Cited on page 188.)

[Mukkamala *et al.* 2013] Raghava Rao Mukkamala, Abid Hussain and Ravi Vatrapu. *Towards a Formal Model of Social Data*. IT University Technical Report Series TR-2013-169, IT University of Copenhagen, Denmark, November 2013. (Cited on pages xvii, 8, 14, 22, 23, 89, 95 and 96.)

[Mukkamala *et al.* 2014] Raghava Rao Mukkamala, Abid Hussain and Ravi Vatrapu. *Towards a Set Theoretical Approach to Big Data Analytics*. In 3rd International Congress on Big Data (IEEE BigData 2014), June 2014. (Cited on pages 8, 22 and 23.)

[Munzner 2014] Tamara Munzner. Visualization analysis and design. CRC Press, 2014. (Cited on page 5.)

[Nam *et al.* 2015] Yoonjae Nam, Yeon-Ok Lee and Han Woo Park. *Measuring web ecology by Facebook, Twitter, blogs and online news: 2012 general election in*

South Korea. Quality & Quantity, vol. 49, no. 2, pages 675–689, 2015. (Cited on page 188.)

[Nash 2008] Jennifer C Nash. *Re-thinking intersectionality*. Feminist review, vol. 89, no. 1, pages 1–15, 2008. (Cited on page 9.)

[National Research Council *et al.* 2013] National Research Council*et al.* Frontiers in massive data analysis. National Academies Press, 2013. (Cited on pages 1, 83 and 95.)

[Neethu & Rajasree 2013] MS Neethu and R Rajasree. *Sentiment analysis in twitter using machine learning techniques*. In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pages 1–5. IEEE, 2013. (Cited on page 26.)

[Nunamaker *et al.* 2017] Jay F Nunamaker, Nathan W Twyman, Justin Scott Giboney and Robert O Briggs. *Creating High-Value Real-World Impact through Systematic Programs of Research*. MIS Quarterly, vol. 41, no. 2, 2017. (Cited on page 20.)

[Ohsumi 2000] Noboru Ohsumi. *From Data Analysis to Data Science*. In Data Analysis, Classification, and Related Methods, pages 329–334. Springer Berlin Heidelberg, 2000. (Cited on page 8.)

[Olshannikova *et al.* 2017] Ekaterina Olshannikova, Thomas Olsson, Jukka Huhtamäki and Hannu Kärkkäinen. *Conceptualizing big social data*. Journal of Big Data, vol. 4, no. 1, page 3, 2017. (Cited on page 2.)

[Pääkkönen 2016] Pekka Pääkkönen. *Feasibility analysis of AsterixDB and Spark streaming with Cassandra for stream-based processing*. Journal of Big Data, vol. 3, no. 1, page 6, 2016. (Cited on page 185.)

[Padmanabhan *et al.* 2014] Anand Padmanabhan, Shaowen Wang, Guofeng Cao, Myunghwa Hwang, Zhenhua Zhang, Yizhao Gao, Kiumars Soltani and Yan Liu. *FluMapper: A cyberGIS application for interactive analysis of massive location-based social media*. Concurrency and Computation: Practice and Experience, vol. 26, no. 13, pages 2253–2265, 2014. CPE-13-0348.R2. (Cited on page 183.)

[Perez 2018] Sarah Perez. *Twitter's doubling of character count from 140 to 280 had little impact on length of tweets*. Techcrunch, December 2018. (Cited on page 6.)

[Pfeffer *et al.* 2015] Karin Pfeffer, Hebe Verrest and Ate Poorthuis. *Big Data for Better Urban Life?–An Exploratory Study of Critical Urban Issues in Two Caribbean Cities: Paramaribo (Suriname) and Port of Spain (Trinidad and Tobago)*. The European Journal of Development Research, vol. 27, no. 4, pages 505–522, 2015. (Cited on page 186.)

[Quinn *et al.* 2016] Martin Quinn, Theodore Lynn, Stephen Jollands and Binesh Nair. *Domestic Water Charges in Ireland-Issues and Challenges Conveyed through Social Media*. Water Resources Management, pages 1–15, 2016. (Cited on page 184.)

[Ramanathan *et al.* 2013] Arvind Ramanathan, Laura L Pullum, Chad A Steed, Chakra Chennubhotla, Shannon Quinn and Tara L Parker. *Oak Ridge Bio-surveillance Toolkit (ORBiT): Integrating Big-Data Analytics with Visual Analysis for Public Health Dynamics*. Technical report, Oak Ridge National Laboratory (ORNL), 2013. (Cited on page 185.)

[Ribarsky *et al.* 2014] William Ribarsky, Derek Xiaoyu Wang and Wenwen Dou. *Social media analytics for competitive advantage*. Computers & Graphics, vol. 38, pages 328 – 331, 2014. (Cited on page 183.)

[Rodgers *et al.* 2015] Peter Rodgers, Gem Stapleton and Peter Chapman. *Visualizing sets with linear diagrams*. ACM Transactions on Computer-Human Interaction (TOCHI), vol. 22, no. 6, page 27, 2015. (Cited on pages xvii, 43, 46 and 47.)

[Ruskey & Weston 1997] Frank Ruskey and Mark Weston. *A survey of Venn diagrams*. Electronic Journal of Combinatorics, vol. 4, page 3, 1997. (Cited on page 43.)

[Scholtz 2004] Jean Scholtz. *Usability evaluation*. National Institute of Standards and Technology, vol. 1, 2004. (Cited on page 34.)

[See-To & Ngai 2016] Eric WK See-To and Eric WT Ngai. *Customer reviews for demand distribution and sales nowcasting: a big data approach*. Annals of Operations Research, pages 1–17, 2016. (Cited on page 190.)

[Sein *et al.* 2011] Maung Sein, Ola Henfridsson, Sandeep Purao, Matti Rossi and Rikard Lindgren. *Action Design Research*. MIS Quarterly, vol. 35, no. 1, pages 37 – 56, 2011. (Cited on pages xvii, 17, 18 and 19.)

[Seo *et al.* 2013] J. Seo, S. Guo and M. S. Lam. *SocialLite: Datalog extensions for efficient social network analysis*. In 2013 IEEE 29th International Conference on Data Engineering (ICDE), pages 278–289, April 2013. (Cited on page 39.)

[Spacey 2018] John Spacey. *5 Types of Design Objectives*, June 2018. (Cited on pages 34 and 53.)

[Sponder 2012] Marshall Sponder. Social media analytics: effective tools for building, intrepreting, and using metrics. McGraw-Hill, 2012. (Cited on page 8.)

[Sterne 2010] Jim Sterne. Social media metrics: How to measure and optimize your marketing investment. John Wiley & Sons, 2010. (Cited on page 8.)

[Subramanian *et al.* 1999] Muralidhar Subramanian, Vishu Krishnamurthy and Redwood Shores. *Performance challenges in object-relational DBMSs*. IEEE Data Eng. Bull., vol. 22, no. 2, pages 27–31, 1999. (Cited on page 54.)

[Sun *et al.* 2013] Guo-Dao Sun, Ying-Cai Wu, Rong-Hua Liang and Shi-Xia Liu. *A survey of visual analytics techniques and applications: State-of-the-art research and future challenges*. Journal of Computer Science and Technology, vol. 28, no. 5, pages 852–867, 2013. (Cited on page 191.)

[Suthers & Rosen 2011] Daniel Suthers and Devan Rosen. *A Unified Framework for Multi-level Analysis of Distributed Learning*. In Proceedings of the 1st International Conference on Learning Analytics and Knowledge, LAK '11, pages 64–74, New York, NY, USA, 2011. ACM. (Cited on pages 22 and 26.)

[Suthers *et al.* 2010] Daniel Suthers, Nathan Dwyer, Richard Medina and Ravi Vatrapu. *A framework for conceptualizing, representing, and analyzing distributed interaction*. International Journal of Computer-Supported Collaborative Learning, vol. 5, no. 1, pages 5–42, 2010. (Cited on pages 22 and 26.)

[Suthers 2017] Daniel Suthers. *Applications of Cohesive Subgraph Detection Algorithms to Analyzing Socio-Technical Networks*. 01 2017. (Cited on page 7.)

[Tichy *et al.* 1979] Noel M Tichy, Michael L Tushman and Charles Fombrun. *Social network analysis for organizations*. The Academy of Management Review, vol. 4, no. 4, October 1979. (Cited on page 7.)

[Tilkov & Vinoski 2010] Stefan Tilkov and Steve Vinoski. *Node. js: Using JavaScript to build high-performance network programs*. IEEE Internet Computing, vol. 14, no. 6, pages 80–83, 2010. (Cited on page 54.)

[Tufekci 2014] Zeynep Tufekci. *Big questions for social media big data: Representativeness, validity and other methodological pitfalls*. arXiv preprint arXiv:1403.7400, 2014. (Cited on page 9.)

[Vaishnavi & Kuechler 2004] Vijay Vaishnavi and William Kuechler. *Design research in information systems*. 2004. (Cited on page 17.)

[Van Deursen *et al.* 2000] Arie Van Deursen, Paul Klint and Joost Visser. *Domain-specific languages: An annotated bibliography*. ACM Sigplan Notices, vol. 35, no. 6, pages 26–36, 2000. (Cited on page 39.)

[Van Welie & Trætteberg 2000] Martijn Van Welie and Hallvard Trætteberg. *Interaction patterns in user interfaces*. In 7th. Pattern Languages of Programs Conference, pages 13–16, 2000. (Cited on page 34.)

[Vatrapu *et al.*] Ravi Vatrapu, Hannu Kärkkäinen, Raghava Rao Mukkamala, Karan Menon, Jukka Huhtamäki, Jari Jussila, **Benjamin Flesch** and Niels Buus Lassen. *Big Social Data Analytics: Past, Present, and Future*. Unpublished Manuscript. (Cited on pages 2, 5 and 15.)

[Vatrapu *et al.* 2014] Ravi Vatrapu, Raghava Rao Mukkamala and Abid Hussain. *A Set Theoretical Approach to Big Social Data Analytics: Concepts, Methods,*

Tools, and Findings. In ECCS Satellite Workshop 2014, pages 22–24, 2014. (Cited on pages 8 and 27.)

[Vatrapu *et al.* 2015] Ravi Vatrapu, Abid Hussain, Niels Buus Lassen, Raghava Rao Mukkamala, **Benjamin Flesch** and Rene Madsen. *Social Set Analysis: Four Demonstrative Case Studies*. In Proceedings of the 2015 International Conference on Social Media & Society, page 3. ACM, 2015. (Cited on pages xvii, 14, 30, 43 and 44.)

[Vatrapu *et al.* 2016] Ravi Vatrapu, Raghava Rao Mukkamala, Abid Hussain and **Benjamin Flesch**. *Social Set Analysis: A Set Theoretical Approach to Big Data Analytics*. IEEE Access: Special Section on Theoretical Foundations for Big Data Applications: Challenges and Opportunities, vol. 4, pages 2542–2571, 2016. (Cited on pages xvii, 2, 4, 8, 9, 13, 20, 22, 23, 24, 83, 84, 95, 96 and 115.)

[Vatrapu 2010] Ravi K. Vatrapu. *Explaining Culture: An Outline of a Theory of Socio-technical Interactions*. In Proceedings of the 3rd International Conference on Intercultural Collaboration, ICIC '10, pages 111–120, New York, NY, USA, 2010. ACM, ACM. (Cited on pages 22 and 26.)

[Vatrapu 2013] Ravi Vatrapu. *Understanding Social Business*. In Emerging Dimensions of Technology Management, pages 147–158. Springer, 2013. (Cited on page 8.)

[Venn 1880] J. Venn. *I. On the diagrammatic and mechanical representation of propositions and reasonings*. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 10, no. 59, pages 1–18, 1880. (Cited on page 43.)

[Viavant *et al.* 2002] Steven Viavant, Arsalan Farooq, Jaydeep Marfatia and Manu Shukla. *Techniques for server-controlled measurement of client-side performance*, December 5 2002. US Patent App. 09/945,160. (Cited on page 53.)

[Vorvoreanu *et al.* 2013] Mihaela Vorvoreanu, Geovon A Boisvenue, Clifford J Wojtalewicz and Eric J Dietz. *Social media marketing analytics: A case study of the public's perception of Indianapolis as Super Bowl XLVI host city*. Journal of Direct, Data and Digital Marketing Practice, vol. 14, no. 4, pages 321–328, 2013. (Cited on page 188.)

[Wamba *et al.* 2017] Samuel Fosso Wamba, Angappa Gunasekaran, Shahriar Akter, Steven Ji-fan Ren, Rameshwar Dubey and Stephen J Childe. *Big data analytics and firm performance: Effects of dynamic capabilities*. Journal of Business Research, vol. 70, pages 356–365, 2017. (Cited on page 1.)

[Ware 2004] Colin Ware. Information Visualization: Perception for Design. Elsevier, San Francisco, CA, USA, 2 édition, April 2004. (Cited on page 5.)

[Wasserman & Faust 1994] Stanley Wasserman and Katherine Faust. Social network analysis: Methods and applications (vol. 8). Cambridge university press, New York, NY, USA, 1 édition, November 1994. (Cited on page 7.)

[Wei *et al.* 2016] Wei Wei, Kenneth Joseph, Huan Liu and Kathleen M Carley. *Exploring characteristics of suspended users and network stability on Twitter*. Social Network Analysis and Mining, vol. 6, no. 1, page 51, 2016. (Cited on pages 9 and 186.)

[Xu *et al.* 2016] Z. Xu, Y. Liu, N. Yen, L. Mei, X. Luo, X. Wei and C. Hu. *Crowdsourcing based Description of Urban Emergency Events using Social Media Big Data*. IEEE Transactions on Cloud Computing, vol. PP, no. 99, pages 1–1, 2016. (Cited on page 186.)

[Yang *et al.* 2016] Jiue-An Yang, Ming-Hsiang Tsou, Chin-Te Jung, Christopher Allen, Brian H Spitzberg, Jean Mark Gawron and Su-Yeon Han. *Social media analytics and research testbed (SMART): Exploring spatiotemporal patterns of human dynamics with geo-targeted social media messages*. Big Data & Society, vol. 3, no. 1, page 2053951716652914, 2016. (Cited on page 183.)

[Yeon *et al.* 2016] Hanbyul Yeon, Seokyeon Kim and Yun Jang. *Predictive visual analytics of event evolution for user-created context*. Journal of Visualization, pages 1–16, 2016. (Cited on page 182.)

[Zimmerman *et al.* 2014] Chris Zimmerman, Yuran Chen, Daniel Hardt and Ravi Vatrapu. *Marius, the giraffe: a comparative informatics case study of linguistic features of the social media discourse*. In Procs. of conference on Collaboration across boundaries: culture, distance & technology, pages 131–140. ACM, 2014. (Cited on page 182.)

Publication I

Social Set Analysis: A Set Theoretical Approach to Big Data Analytics

---

Ravi Vatrapu, Raghava Rao Mukkamala, Abid Hussain and **Benjamin Flesch**. *Social Set Analysis: A Set Theoretical Approach to Big Data Analytics*. IEEE Access: Special Section on Theoretical Foundations for Big Data Applications: Challenges and Opportunities, vol. 4, pages 2542–2571, 2016

Social Set Analysis: A Set Theoretical Approach to Big Data Analytics

Ravi Vatrapu[1,2], Raghava Rao Mukkamala[1], Abid Hussain[1] and Benjamin Flesch[1]
[1]Computational Social Science Laboratory (http://cssl.cbs.dk), Copenhagen Business School, Denmark and
[2]Westerdals Oslo School of Arts, Comm & Tech, Norway
{rv.itm, rrm.itm, ah.itm, bf.itm}@cbs.dk

***Abstract*—Current analytical approaches in Computational Social Science can be characterized by four dominant paradigms: text analysis (information extraction and classification), social network analysis (graph theory), social complexity analysis (complex systems science), social simulations (cellular automata and agent-based modelling). However, when it comes to organizational and societal units of analysis, there exists no approach to conceptualise, model, analyze, explain and predict social media interactions as individuals' associations with ideas, values, identities, etc. To address this limitation, based on the sociology of associations and the mathematics of set theory, this paper presents a new approach to big data analytics called Social Set Analysis. Social Set Analysis consists of a generative framework for philosophies of computational social science, theory of social data, conceptual and formal models of social data, and an analytical framework for combining big social datasets with organisational and societal datasets. Three empirical studies of big social data are presented to illustrate and demonstrate Social Set Analysis in terms of fuzzy set-theoretical sentiment analysis, crisp set-theoretical interaction analysis and event-studies oriented set-theoretical visualisations. Implications for big data analytics, current limitations of the set-theoretical approach, and future directions are outlined.**

***Index Terms*—Big social data, Formal Models, Social Set Analysis, Big data visual Analytics, New Computational Models for Big Social Data.**

I. Introduction

Social media are fundamentally scalable communications technologies that turn Internet based communications into an interactive dialogue platform [1]. On the "demand-side", users and consumers are increasingly turning to various types of social media to search for information and to make decisions regarding products, politicians, and public services [2]. On the "supply-side", terms such as "Enterprise 2.0" [3] and "social business" [4] are being used to describe the emergence of private enterprises and public institutions that strategically adopt and use social media channels to increase organizational effectiveness, enhance operational efficiencies, empower employees, and co-create with stakeholders. The organizational and societal adoption and use of social media is generating large volumes of unstructured data that is termed *Big Social Data*. New organizational roles such as Social Media Manager, Chief Listening Officer, Chief Digital Officer, and Chief Data Scientist have emerged to meet the associated technological developments, organizational changes, market demands, and societal transformations. However, the current state of knowledge and practice regarding social media engagement is rife with numerous technological problems, scientific questions, operational issues, managerial challenges, and training deficiencies. As such, not many organizations are generating competitive advantages by extracting meaningful facts, actionable insights and valuable outcomes from Big Social Data analytics. Moreover, there are critical unsolved problems regarding how Big Social Data integrates with the existing datasets of an organization (that is, data from internal enterprise systems) and its relevance to the organisation's key performance indicators. To address these diverse but interrelated issues, this paper presents a novel set-theoretical approach to Big Data Analytics in general and Big Social Data Analytics in particular for Facebook, Twitter and other social media channels.

Specifically, this paper introduces a research program situated in the domains of Data Science [5]–[7] and Computational Social Science [8] with practical applications to Social Media Analytics in organizations [4], [9], [10]. It addresses some of the important theoretical and methodological limitations in the emerging paradigm of Big Data Analytics of social media data [11]. From an academic research standpoint, Social Set Analysis addresses two major limitations with the current state of the art in Computational Social Science: (i) a vast majority of the extant literature is on twitter datasets with only 5% of the papers analysing Facebook data raising representativeness, validity and methodological concerns [11], and (ii) mathematical modelling of social data hasn't progressed beyond the four dominant approaches [12] of text analysis (information extraction and classification), social network analysis (graph theory), social complexity analysis (complex systems science), social simulations (cellular automata and agent-based modelling).

To put it honestly and provocatively, currently we don't have deep academic knowledge of the most dominant action on social media platforms performed by hundreds of millions of unique users every day: "like" on Facebook. In fact, as Claudio Cioffi-Revilla (2013), one of the founding parents of the field of Computational Social Science, astutely observed:

> *Reliance on the same mathematical structure every time (e.g., game theory, as an example), for every research problem, is unfortunately a somewhat common methodological pathology that leads to theoretical decline and a sort of inbreeding visible in some areas of social science research. Dimensional empirical features of social phenomena-such as discreteness-continuity, deterministic-stochastic, fnite-infnite, contiguous-isolated, local-global, long-term vs. short-term, independence-interdependence,*

synchronic-diachronic, among others-should determine the choice of mathematical structure(s).

This lack of mathematical imagination coupled with hyperactive boundary-policing of the "purity of the turf" of Computational Social Science results in major conceptual and technical limitations when analysing big social data resulting from individuals' and organizations' Facebook and Twitter engagement. There is both a research gap and real-world organisational needs to describe, model, analyse, explain, and predict such interactions as individuals' associations to ideas, values, identities etc [13].

For example, a typical post on F.C. Barcelona's Facebook page generates around 100,000 unique likes, 5,000 comments and 1,000 shares). Facebook users' "likes" on any given F.C. Barcelona post could be personal-association to one of the players, identity-association to the Catalan, political-association to pro-independence parties of Catalonia, brand-association to the corporate sponsors etc. The mathematics of set theory is ideally suited to model such associations in the first analysis. Just like graph theory is ideally suited for Social Network Analysis [14] of dyadic relations from the perspective of relational sociology [15], set theory is ideally suited for conceptualising, modelling, and analysing monadic, dyadic, and polyadic human associations to ideas, values and identities [16] from the perspective of the sociology of associations. This is the gist of the set theoretical approach proposed by this paper.

A. Overarching Research Question

In order to further research in this area we as ourselves the following research question:
How can models, methods and tools for Social Set Analysis derived from the alternative holistic approach to Big Social Data Analytics based on the sociology of associations and the mathematics of set theory result in meaningful facts, actionable insights and valuable outcomes?

II. Conceptual Framework

A. Need for a Philosophy of Computational Social Science

The purpose of this section is to present an argument that we need philosophies of Computational Social Science that explicilty outline and discuss their sociological assumptions, mathematical modelling, computational implementation, and empirical analysis. To the best of our knowledge, no such philosophy of Computational Social Science exists other than Social Network Analysis [17] based on the mathematics of graph theory [18] and the sociology of relations [15]. However, the philosophical assumptions of relational sociology might be not be relevant to all classes of problems in computational social science. For example, for the class of problems that address big social data from the Facebook or Twitter interactions of large brands such as Coca-Cola or a F.C. Barcelona, the fundamental assumption of SNA that social reality is constituted by dyadic relations and interactions are determined by structural positions of individuals in social networks [19] is neither necessary nor sufficient [20]. Other dominant paradigms of computational social science such as Social Complexity and Social Simulation [12] have varying levels of philosophical and modelling unity and maturity. [12]. Therefore, there is a clear need for a manifest statement and critical examination of philosophical principles that underpin the theoretical, methodological, and analytical aspects of current Computational Social Science approaches.

However, philosophical proposals for Big Data Analytics must avoid the malaise of *over-philosophising* with non-realist ontologies and non-empirical epistemologies (for a precautionary tale from the Humanities and Social Sciences, please cf. [21], [22]) that result in little-to-no methodological innovation in terms of instrumentation, measurement and evaluation of the phenomena of interest. Philosophical framworks for Big Data Analytics should aspire towards positive contributions that go beyond the negative criticisms of assumptions and methods that regularly feature in prominent recent criticisms (for instance, [11], [23]). We argue that one class of positive contributions would be generative frameworks that provide explicit articulation of philosophical assumptions underlying analytical approaches as well as a *production system* for creating and evaluating new philosophies. To address the analytical limitations identified and to fulfill the critical and generative criteria outlined above, we propose a first version of the generative framework for the philosophy of Computational Social Science.

1) A Generative Framework for Philosophy of Computational Social Science (GF-PCSS): The preliminary version of the GF-PCSS comprising of five elements is presented in Table I below.

Philosophical Dimension	GF-PCSS Element	Key Assumptions
Ontology	Basic Premise	What is social?
		When is it social?
		Being vs. Becoming of social
Epistemology	Social Action	How is it social?
		How does a social entity act and interact?
Methodological	Unit of Analysis	What is the foundational analytical unit?
		What is the minimum viable analytical entity?
Political	Social Structure	What is the social grouping entity?
		What is the social formation unit?
Formal	Mathematics	What is the appropriate mathematical theory for modelling?

Table I
Five Elements of the Candidate Generative Framework for Philosophy of Computational Social Science

Given the preliminary stage of the GF-PCSS, no claims are made about the exhaustiveness and/or mutual exclusivity of the five elements. We simply claim that the five elements are necessary with no claims made about their sufficiency and orthogonality.

Table II below seeks to illustrate the positive contribution of the GF-PCSS. First, the framework is used to explicitly

state the latent philosophical assumptions of one dominant traditional approach in Computational Social Science, Social Network Analysis. Second, the framework is used to better understand the limitations of Social Network Analysis with respect to large-scale social media platforms that are increasingly content driven. Social Network Analysis is primarily concerned with how social actors relate to each other and not so much with how content is generated, interacted and circulated in terms of ideas, aspirations, values, and identities. However, large-scale and content driven social media platforms such as Facebook are of extreme importance to organizations in terms of marketing communications, corporate social responsibility, democratic deliberation, public dissemination etc. Social media analytics in practice [9], [10], [24] has been based on an implicit, inherent and latent understanding of human associations as expressed by metrics and key performance indicators such as brand sentiment, brand associations, conversation keywords, reach etc. Further, Social Network Analysis assumes *homophiliy* rather than explaining the agentic mechanisms constituting it. Third and last, GF-PCSS is used to generate a new holistic approach termed Social Set Analysis and make a positive contribution. Social Set Analysis is based on the philosophical principles derived from ecological psychology, micro sociology, associational sociology [25], and the mathematics of the set theory (crisp sets, fuzzy sets, rough sets, and random sets) [26].

	Social Network Analysis	**Social Set Analysis**
Basic Premise	There exists a **relation** between social actor A and social actor B	There exists an **association** by actor A with some entity E which can be an actor or an artifact
Social Action	Molecular Relations	Atomic Actions
Unit of Analysis	Dyadic	Monadic, Dyadic & Polyadic
Social Configuration	Networks	Sets
Social Explanation	Structural	Agentic
Mathematics	Graph Theory	Set Theory

Table II
CONTRASTING PHILOSOPHIES OF COMPUTATIONAL SOCIAL SCIENCE

To be clear, our argument is not that current approaches in Computational Social Science such as Social Network Analysis (based on relational sociology, graph theory, and network analysis) are invalid or ineffective. Instead, our argument, as articulated and illustrated in Tables I & II, is that a generative framework of the philosophy can be used to make a fundamental change in the foundational mathematical logic of the formal model from graphs to sets which can yield new analytical insights for a new class of problems (in our case, organizational use of social media).

B. Set Theoretical Big Social Data Analytics

As articulated in [27], based on Smithson and Verkuilen [28] there are five advantages to applying classical set theory [29] in general and fuzzy set theory [26] in particular to computational social sciences:

1) Set-theoretical ontology is well suited to conceptualize vagueness, which is a central aspect of social science constructs. For example, in the social science domain of marketing, concepts such as brand loyalty, brand sentiment and customer satisfaction are vague.
2) Set-theoretical epistemology is well suited for analysis of social science constructs that are both categorical and dimensional. That is, set-theoretical approach is well suited for dealing with different and degrees of a particular type on construct. For example, social science constructs such as culture, personality, and emotion are all both categorical and dimensional. A set-theoretical approach can help conceptualize their inherent duality.
3) Set-theoretical methodology can help analyze multivariate associations beyond the conditional means and the general linear model. In addition, set theoretical approaches analyze human associations prior to relations and this allows for both quantitative variable centered analytical methods as well as qualitative case study methods.
4) Set-theoretical analysis has high theoretical fidelity with most social science theories, which are usually expressed logically in set-terms. For example, theories on market segmentation and political preferences are logically articulated as categorical inclusions and exclusions that natively lend themselves to set theoretical formalization and analytics.
5) Set-theoretical approach systematically combines set-wise logical formulation of social science theories and empirical analysis using statistical models for continuous variables. For example, in the case of predictive analytics, it is possible to employ set and fuzzy theory to dynamically construct data points for independent variables such as brand sentiment (polarity, subjectivity, etc.).

We now present a theory of social data based on the philosophical framework for Social Set Analysis discussed above.

C. Theory of Social Data

For the purposes of systematically collecting and analysing big social data, we argue that any candidate theory of social data must support conceptual and mathematical modelling of data at the software log level. After all, it is a fact that the outcomes from big social data collection from modern web service calls or historic web crawling methods are nothing more than digital trace records and software log entries. As such, an appropriate theory of social data would be operational at the micro-genetic level of social media interactions as they unfold in the real-time and in the actual-space of a computer screen of some kind (desktop monitor, laptop display or the mobile phone screen). For Social Set Analysis, we have selected the theory of socio-technical interactions by Vatrapu [30]–[32] as it conceptualises perception of and interaction on the screen in real-time and actual-space. The theory of socio-technical interations [30]–[32] is derived from the following sources:

1) the ecological approach to perception and action [33]
2) the enactive approach to the philosophy of mind [34]

3) the phenomenological approach to sociology [35], [36]

A more detailed exposition of the theory of socio-technical interactions regarding its ontological and epistemological assumptions and principles, is beyond the scope of this paper but for a consise overview, please confer [32].

We use the theory of socio-technical interactions [30]–[32] to describe how individual data items (or trace records) such as Facebook posts, likes, comments etc. come into being. In other words, we use the theory of socio-technical interactions to describe the phenomenon of big social data generation from its constituent individual interactions of Facebook posts, comments, likes etc. That said, the scope and extent of the theory of social data are restricted to providing phenomenological grounding for modelling of the social data retrieved from social media platforms such as Facebook. The theory of social data is outlined the next subsection (II-D).

As already mentioned, the theory of social data is drawn from the theory of socio-technical interactions [30]–[32]. Social media platforms such as Facebook and Twitter, at the highest level of abstraction, involve individuals interacting with (a) technologies and (b) other individuals. These interactions are termed socio-technical interactions and there are two types of socio-technical interactions:

I) Interacting with the technology: An example could be using the Facebook app on the user's smartphone.
II) Interacting with others socially using the technology: An example could be liking a picture posted by a friend in the Facebook app on the user's smartphone.

These socio-technical interactions are theoretically conceived as

I) Perception and appropriation of socio-technical affordances
II) Structures and functions of technological intersubjectivity

Briefly, socio-technical affordances are action-taking possibilities and meaning-making opportunities in an actor-environment system bound by the cultural-cognitive competencies of the actor and the technical capabilities of the environment. Technological intersubjectivity (TI) [30]–[32] refers to a technology supported, interactional social relationship between two or more actors.

Socio-technical interactions as described above result in electronic trace data that is termed *social data*. In case of the previously mentioned example where a Facebook user liking a picture posted by a friend on their smartphone app, the social data is not only rendered in the different *timelines* of the user's social network but it is available via the Facebook graph API. Large volumes of such micro-interactions constitute the macro world of Big Social Data which is the analytical focus of this paper. Our argument is not that there exists only one set of candidates for the theory of social data, conceptual model of social data and the formal model of social data as proposed in this paper. Instead, our argument is that a theoretically informed and empirically oriented research project in big social data analytics must incoporate these components (theory, conceptual and formal models of social data) and computationally realise each of them within IT-Artefacts.

D. Conceptual Model of Social Data

Based on the theory of social data described above, we present the conceptual model of social data below.

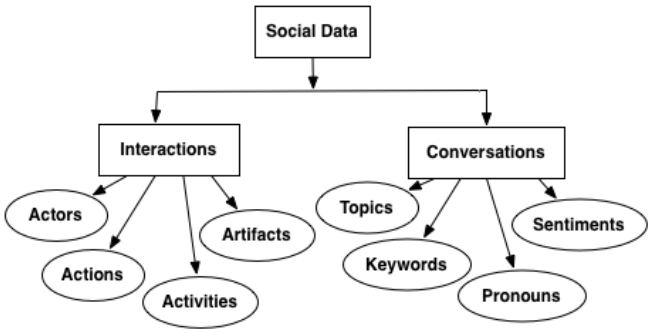


Figure 1. Conceptual Model of Social Data [37]

In general, *Social data* consists of two types: (a)Interactions (what is being done) and Conversations (what is being said). Interactions refer to the first aspect of socio-technical interactions constituted by the perception and appropriation of affordances (which users/actors perceive which socio-techncial affordances to interact with what other social actors on social media platforms). Conversations relates to the second aspect of socio-technical interactions: structures and functions of technological intersubjectivity (what the users/actors are communicating to each other and how they are influencing each other through both natural language as well as design language of the social media plaforms). Interactions consists of the structure of the relationships emerging from the appropriation of social media affordances such as posting, linking, tagging, sharing, liking etc. It focuses on identifying the actors involved, the actions they take, the activities they undertake, and the artifacts they create and interact with. Conversations consists of the communicative and linguistic aspects of the social media interaction such as the topics discussed, keywords mentioned, pronouns used and emotions expressed. Figure 1 presents the conceptual model of social data.

E. Illustrative Example of Social Data

Let us say that the research domain is Corporate Social Responsbility (CSR) and the research question is to what extent do Facebook walls function as online public spheres with regard to CSR in terms of marketing campaigns as well as crises. Then, the set theoretical approach to computational social science can be employed to specify measures of the extent to which the Facebook Walls are serving as online public spheres as discussed below.

Focusing on the interactional aspect of big social data allows the examination of the breadth of engagement of the public sphere by reporting the overall number of posts made (artifacts), which of the Facebook walls received most posts and whether they linked out to other sources of information. In addition to looking at the posts in the aggregate we also can look at them individually and map linkage across walls. Was the posting entirely independent in that individuals (actors) only posted (action) to one wall or did they post more widely

on two or three walls' Interactional analysis of big social data can help reveal the patterns and dynamics of actors' mobility across space (multiple facebook walls) and time (before, during and after campaigns/crises).

Focusing on the conversational aspect of big social data allows the examination of the depth of the engagement taking place through the Facebook walls and thus whether walls are acting as an online public space. In particular we can look at four key aspects of the posts and comments: topics, keywords and emotions. As with interactional analysis, conversational analysis of big social data can help reveal the patterns and dynamics of actors' conversational genres across space (multiple facebook walls) and time (before, during and after campaigns/crises).

III. Research Methodology

Our research methodology is shown in Fig. 2 and is described below:

1) Systematically collect big social data about organisations from Facebook, Twitter etc using the Social Data Analytics Tool [37], [38] developed in the Computational Social Science Laboratory (http://cssl.cbs.dk) and other research and commercial tools.
2) Technically combine organisational process data with business social data so that the resulting dataset legally compliant, ethically correct, privacy adherent, and data security ensured
3) Big Social Data Analytics: Phase One: Adopt current methods, techniques and tools from Computational Social Science to model and analyse
   a) Interaction Analysis: Social Network Analysis, Complex Systems Dynamics, Event Study Methodology from Finance, Data Mining from Computer Science
      - Who is doing what, when, where, how and with whom?
      - Social media users and organisational stakeholders (like consumers) liking pictures of cute puppies posted by Walmart on its official Facebook wall every third Sunday according to its social media marketing strategy.
   b) Conversation Analysis: Computational Linguistics & Machine Learning
      - What are the things human actors (and fraudulent accounts/robots) saying?
      - Social media users and organisational stakeholders (like consumers) commenting on those pictures of cute puppies by discussing/mentioning various topics/keywords of organisational/societal relevance/irrelevance and expressing their subjective feelings etc.
4) Applying set theoretical methods and techniques drawn from crisp sets, fuzzy sets, rough sets and random sets [26], [28], [29], [39].
5) Software realisation of the empirical findings from traditional and novel (set theoretical) approaches to Computational Social Science as a tools for Organisations.
6) Publication of research findings in peer-reviewed conferences, journals and edited books.
7) Generation of instrumental benefits for Organisations in terms of meaningful facts (sensible data), actionable insights (applicable information), valuable outcomes (constructive knowledge) and sustainable impacts (wisdom)

IV. Related Work

A. Social Network Analysis

Social Network Analysis can be traced back to 1979, where Tichy et.al. [40] used it as a method to examine the relationships and organisational social structures. In the later years, cognitive social structures as a solution for social network related problems was proposed by David Krackhardt [41]. The field of social computing attracted many researchers due to the latest developments of online social media since last decade. Even though It is not possible to refer to an extensive list of research articles in this emerging area, however we refer some of the latest and important works here.

In their research article, Justin Zhan and Xing Fang in [42] provided an detailed overview about research on social networking analysis, human behavioural modelling and security aspects in the context of social networks. Social network analysis based on measuring social relations using multiple data sets has been explored in [43]. In the context of multi-agent systems using social network analysis, a framework for calculating reputations has been proposed by [44]. An algorithm to find overlapping communities via social network analysis was explored in [45]. Moreover, analysis of subgraphs in the social network based on the characteristic features: leadership, bonding, and diversity was studied by the authors in [46]. All these works focussed on using social network analysis and other graph related formalisms as main tools for analysis of social media where the primary focus is on the structural aspects of social data. On the other hand, our work primarily focussed on using set theory and fuzzy logics for analysis of both structure and content of social media data. Therefore we are not only interested in analyzing the structural aspects of social data (as networks or sets) but also in understanding the substantive aspects of social data (as sentiments, topics, keywords, pronouns).

B. Social Text Analysis

A comprehensive state-of-the-art review of computational linguistics is provided by Pang and Lee [47]. They provided approaches to analyse natural language texts, and identify three different technical terms: opinions, sentiments, and subjectivity. In this paper, we adopt Pang and Lee's [47] technical interpretation that opinion mining and sentiment analysis can be treated as identical and conduct sentence level rather than sub-sentence level sentiment analysis as discussed in [48]. Other methods and techniques for sentiment analysis are presented and discussed in [47]–[52]. Below is a selected listing of related work in sentiment analysis of social data ranging over a variety of methods, techniques, and tools.

Prior work has shown sentiment analysis of social data can be used to predict movie revenues [53], correlate with contemporaneous and subsequent stock returns [54], exploring cultural and linguistic differences in ratings and reviews [55],

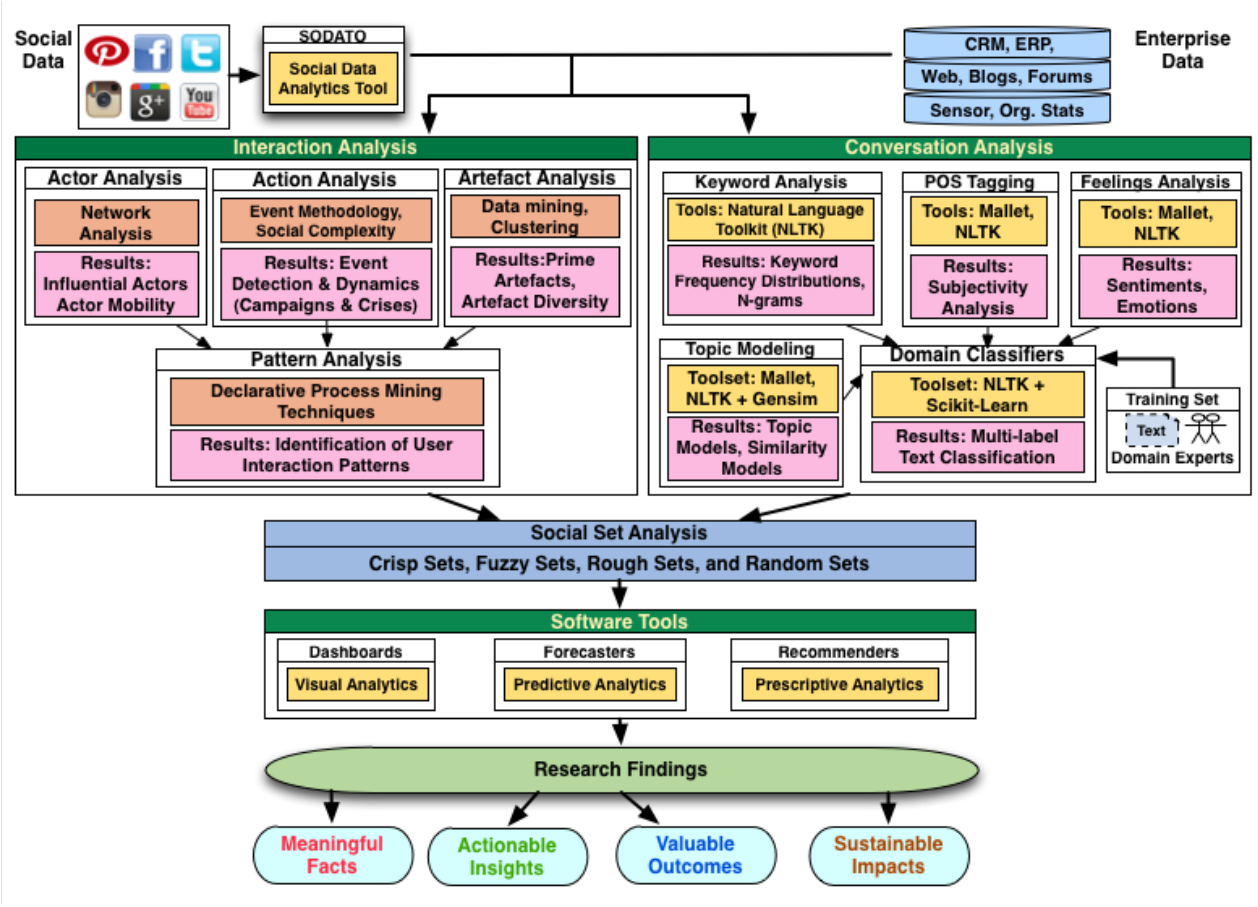


Figure 2. Research Framework for Set-Theoretical Big Social Data Analytics

sentiment evolution in political deliberation on social media channels [56], assess sentiment towards a new vaccine [57], and explore semantic-level precedence relationships between participants in a blog network [58]. To briefly expand, [58] proposed a methodology for the detection of bursts of activity at the semantic level using linguistic tagging, term filtering and term merging, where a probabilistic approach was used to estimate temporal relationships between the blogs. Asur and Huberman [53] showed that sentiment analysis on Twitter's content urls, retweets and their hourly rates can predict box-office movies revenues to a high degree of precision.

In contrast to the existing approaches, we used Set and Fuzzy Set Theory for the formal modelling of associations between actors, actions, artifacts, topics and sentiments in order to provide a systemic treatment of relationship, vagueness and uncertainty in the social data. The existing sentiment analysis techniques (as cited above) use only the classification of individual artifacts (such as either positive or negative or neutral), but not the probabilities associated with the classification labels returned by the sentiment analysis method and/or tool. In contract, our approach uses fuzzy sets to represent artifact sentiment with classification along with their probabilities (e.g. positive: 0.20, negative: 0.65, neutral: 0.15) as explained later.

V. Formal Model of the Conceptual Social Data

In this section, we will provide formal semantics for the concepts of social data, which is based on social data model that was initially presented in [27], [59], but refined according to the changes in the conceptual model of social data presented in Sec. II-C.

Notation: For a set A we write $\mathcal{P}(A)$ for the power set of A (i.e. set of all subsets of A) and $\mathcal{P}_{disj}(A)$ for the set of mutually disjoint subsets of A. The cardinality or number of elements in a set A is represented as $\mid A \mid$. Furthermore, we write a relation R from set A to set B as $R \subseteq A \times B$. A function f defined from a set A to set B is written as $f : A \rightarrow B$, where a if f is a partial function then it is written as $f : A \rightharpoonup B$.

First, we define type of artifacts in a socio-technical system as shown in Def. 5.1.

Definition 5.1: We define $\mathbb{R}$ as a set of all artifact types as $\mathbb{R}$ = { status, comment, link, photo, video }.

Definition 5.2: We define $\mathbb{A}_{\mathbb{CT}}$ as a set of actions that can be performed as $\mathbb{A}_{\mathbb{CT}}$ = {*post*, *comment*, *share*, *like*, *tagging*}.

As explained in the conceptual model (Sec. II-C), the *Social Data* model contains *Interactions* (what is being done) and *Conversations* (what is being said). The formally *Social Data* is defined in Def. 5.3 as follows,

Definition 5.3: Formally, Social Data is defined as a tuple $\mathsf{D} = (\mathsf{I}, \mathsf{C})$ where

(i) I is the *Interactions* representing the structural aspects of social data as defined further in Def. 5.4

(ii) C is the *Conversations* representing the content of social data and is further defined in Def. 5.5

Definition 5.4: The Interactions of *Social Data* are defined as a tuple $\mathsf{I} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r_{type}}, \rhd, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$ where

(i) U is a (finite) set of actors (or users) ranged over by u,

(ii) R is a (finite) set of artifacts (or resources) ranged over by r,

(iii) Ac is the activities set which is also finite,

(iv) $\mathsf{r_{type}} : \mathsf{R} \rightarrow \mathbb{R}$ is typing function for artifacts that maps each artifact to an artifact type defined in 5.1,

(v) $\rhd : \mathsf{R} \rightharpoonup \mathsf{R}$ is a partial function mapping artifacts to their parent artifact,

(vi) $\rightarrow_{post} : \mathsf{U} \rightharpoonup \mathcal{P}_{disj}(\mathsf{R})$ is a partial function mapping actors to mutually disjoint subsets of artifacts created by them

(vii) $\rightarrow_{share} \subseteq \mathsf{U} \times \mathsf{R}$ is a relation mapping between users to their artifacts (shared by them),

(viii) $\rightarrow_{like} \subseteq \mathsf{U} \times \mathsf{R}$ is a relation mapping users to the artifacts liked by them,

(ix) $\rightarrow_{tag} \subseteq \mathsf{U} \times \mathsf{R} \times (\mathcal{P}(\mathsf{U} \cup \mathrm{Ke}))$ is a tagging relation mapping artifacts to power sets of actors and keywords indicating tagging of actors and keywords in the artifacts, where Ke is set of keywords defined in Def. 5.5,

(x) $\rightarrow_{\mathsf{act}} \subseteq \mathsf{R} \times \mathrm{Ac}$ is a relation from artifacts to activities.

Formal definition of Interactions is provided in Def. 5.4, where the first three items (i, ii, x of Def. 5.4) contain a set of actors (U), a set of artifacts/resources (R) and a set of activities (Ac). Each artifact is mapped to an artifact type (such as status, photo etc) by artifact type function (Def. 5.4-iv). Furthermore, some of the artifacts are mapped to their parent artifact (if exists) by parent artifact function $\rhd$ (Def. 5.4-v). For example, a comment is an artifact which is made on a post, then it is mapped to its parent (which is the post), on the other hand, if the artifact is a status message or a new post, then there will not be any mapping for that artifact, as it has no parent.

Furthermore, each artifact is posted (created) by a single actor. As shown in Def. 5.4-vi, the $\rightarrow_{post}$ is a partial function mapping actors to mutually disjoint sub sets of artifacts, each set containing artifacts created or posted by an actor. On contrary, the $\rightarrow_{share}$ indicates a many-to-many relationship, indicating that an artifact can be shared by many actors and similarly each actor can share many artifacts (Def. 5.4-vii). Even though *share* and *post* actions seems to be similar, the $\rightarrow_{post}$ signifies the creator relationship of an artifact, where as $\rightarrow_{share}$ indicates share relationship between an artifact and an actor which can be many-to-many.

Similar to the *share* relation, the *like* relation ($\rightarrow_{like}$) maps between the artifacts and actors, indicating the artifacts liked by the actors. The *tagging* relation ($\rightarrow_{tag}$) is a bit different, which is a mapping between actors, artifacts and power set of actors and keywords (Def. 5.4-ix). The basic intuition behind the tag relation is that, it allows an actor to tag other actors or keywords in an artifact. Finally, the $\rightarrow_{\mathsf{act}}$ relation indicates a mapping between artifacts to activities (Def. 5.4-x).

Definition 5.5: In Social Data $\mathsf{D} = (\mathsf{I}, \mathsf{C})$, we define Conversations as C= (To, Ke, Pr, Se, $\rightarrow_{\mathsf{topic}}, \rightarrow_{\mathsf{key}}, \rightarrow_{\mathsf{pro}}, \rightarrow_{\mathsf{sen}}$) where

(i) $\mathrm{To}, \mathrm{Ke}, \mathrm{Pr}, \mathrm{Se}$ are finite sets of topics, keywords, pronouns and sentiments respectively,

(ii) $\rightarrow_{\mathsf{topic}} \subseteq \mathsf{R} \times \mathrm{To}$ is a relation defining mapping between artifacts and topics,

(iii) $\rightarrow_{\mathsf{key}} \subseteq \mathsf{R} \times \mathrm{Ke}$ is a relation mapping artifacts to keywords,

(iv) $\rightarrow_{\mathsf{pro}} \subseteq \mathsf{R} \times \mathrm{Pr}$ is a relation mapping artifacts to pronouns,

(v) $\rightarrow_{\mathsf{sen}} \subseteq \mathsf{R} \times \mathrm{Se}$ is a realtion mapping artifacts to sentiments.

The *Conversations* of Social Data is formally defined in Def. 5.5 and it mainly contains sets of *topics* (To), *keywords* (Ke), *pronouns* (Pr), and *sentiments* (Se) as defined in Def. 5.5. The $\rightarrow_{\mathsf{topic}}$, $\rightarrow_{\mathsf{key}}$, $\rightarrow_{\mathsf{pro}}$ and $\rightarrow_{\mathsf{sen}}$ relations map the artifacts to the *topics* (To), *keywords* (Ke), *pronouns* (Pr), and *sentiments* (Se) respectively. One may note that all these relations allow many-to-many mappings, for example an artifact can be mapped to more than one sentiment and similarly a sentiment can contain mappings to many artifacts.

Finally, we define a time function to record the timestamp of actions performed on social data as follows.

Definition 5.6: In Social Data, let $\mathbf{T} : (u, r, ac) \mapsto \mathbb{N}$ be time function that keeps tracks of timestamp ($t \in \mathbb{N}$) of an action ($ac \in \mathbb{A}_{\mathbb{CT}}$) performed by an actor ($u \in \mathsf{U}$) on an artifact ($r \in \mathsf{R}$).

A. Operational Semantics

Operational semantics of Social Data model are defined in this section. More precisely, we define how actors perform actions on artifacts. As formally defined in Def. 5.7, the first action is *post*, which accepts a pair containing an actor and a new artifact (u, r). First, the actor will be added to the set of actors (i) and then the new artifact will be added to the set of artifacts (ii). Finally the post relation ($\rightarrow_{post}$) will be updated for the new mapping (iii).

Definition 5.7: In Social Data $\mathsf{D} = (\mathsf{I}, \mathsf{C})$ with Interactions $\mathsf{I} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r_{type}}, \rhd, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$, we define a **post** operation of posting a new artifact r ($r \notin \mathsf{R}$) by an user u as $\mathsf{D} \bigoplus_p (u, r) = (\mathsf{I}', \mathsf{C})$ where $\mathsf{I}' = (\mathsf{U}', \mathsf{R}', \mathrm{Ac}, \mathsf{r_{type}}, \rhd, \rightarrow_{post}{}', \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$,

(i) $\mathsf{U}' = \mathsf{U} \cup \{u\}$

(ii) $\mathsf{R}' = \mathsf{R} \cup \{r\}$

(iii) $\rightarrow_{post}{}' = \begin{cases} \rightarrow_{post}(u) \cup \{r\} & \text{if } \rightarrow_{post}(u) \text{ defined} \\ \rightarrow_{post} \cup \{u, \{r\}\} & \text{otherwise} \end{cases}$

The *comment* action (e.g. on a post) accepts a tuple containing an actor, the parent artifact (on which the comment is made) and the comment content itself as shown in the Def. 5.8. As it creates a new artifact, it will first apply a *post* action to create the comment as a new artifact with the actor (i) and then followed by an update to the parent artifact function ($\rhd$) by adding the respective mapping to its parent (ii).

Definition 5.8: In Social Data $\mathsf{D} = (\mathsf{I}, \mathsf{C})$ with Interactions $\mathsf{I} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r_{type}}, \rhd, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$,

the **comment** operation on an artifact r_p ($r_p \in \mathsf{R}$) by an user u for a new artifact r is formally defined as $\mathsf{D} \bigoplus_c (u, r, r_p) = (\mathsf{I}', \mathsf{C})$ where $\mathsf{I}' = (\mathsf{U}', \mathsf{R}', \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright', \rightarrow_{post}{}', \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$,

(i) $\mathsf{D} \bigoplus_p (u, r) = (\mathsf{I}'', \mathsf{C})$ where $\mathsf{I}'' = (\mathsf{U}', \mathsf{R}', \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}{}', \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$,

(ii) $\triangleright' = \triangleright \cup \{r, r_p\}$

As mentioned before, the *share* operation does not create any new artifact, but it will updates the actors set and then makes an update to the share relation ($\rightarrow_{share}$) as formally defined in Def. 5.9.

Definition 5.9: Let Social Data be $\mathsf{D} = (\mathsf{I}, \mathsf{C})$ with Interactions $\mathsf{I} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$, then we define the **share** operation on an artifact r by an user u as $\mathsf{D} \bigoplus_s (u, r) = (\mathsf{I}', \mathsf{C})$ where $\mathsf{I}' = (\mathsf{U} \cup \{u\}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}, \rightarrow_{share} \cup \{(u, r)\}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$.

The following definition (Def. 5.10) contains formal definitions of *like* and *unlike* operations as an update to the like relation ($\rightarrow_{like}$). A *like* action on an artifact will add a mapping to like relation ($\rightarrow_{like}$) (in addition to adding the actor to the actors set), where as an *unlike* action will simply remove the existing mapping.

Definition 5.10: In Social Data $\mathsf{D} = (\mathsf{I}, \mathsf{C})$ with Interactions $\mathsf{I} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$, we define the **like** operation by an user u on an artifact r as $\mathsf{D} \bigoplus_l (u, r) = (\mathsf{I}', \mathsf{C})$ where $\mathsf{I}' = (\mathsf{U} \cup \{u\}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like} \cup \{(u, r)\}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$.

Similarly, we define the **unlike** operation on $\mathsf{D} = (\mathsf{I}, \mathsf{C})$ with Interactions $\mathsf{I} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$, as $\mathsf{D} \bigominus_l (u, r) = (\mathsf{I}', \mathsf{C})$ where $\mathsf{I}' = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like} \setminus \{(u, r)\}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$.

Finally, *tagging* action accepts a tuple ((u, r, t)) containing an actor, an artifact and a set of hash words (i.e. keywords and actors) and an update to tagging relation ($\rightarrow_{tag}$) will be applied as shown in the Def. 5.11.

Definition 5.11: In a Social Data $\mathsf{D} = (\mathsf{I}, \mathsf{C})$ with Interactions $\mathsf{I} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$, we define the **tagging** operation by an user u on an artifact r with a set of hash words $t \in \mathcal{P}(\mathsf{U} \cup \mathrm{Ke})$ as $\mathsf{D} \bigoplus_t (u, r, t) = (\mathsf{I}', \mathsf{C})$ where $\mathsf{I}' = (\mathsf{U} \cup \{u\}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag} \cup \{(u, r, t)\}, \rightarrow_{\mathsf{act}})$.

B. Illustrative Example

In this section, we exemplify the formal model by taking an example post from the Facebook page of McDonald's Food/Beverages as shown in the figure 3. In order to enhance the readability of the example, the artifacts (e.g. texts) have been annotated as $r1, r2$ etc and the annotated values will be used in encoding the example using the formal model.

Example 5.1: The following are some of the texts extracted from a sample post [60] from Facebook page of McDonald's Food/Beverages.

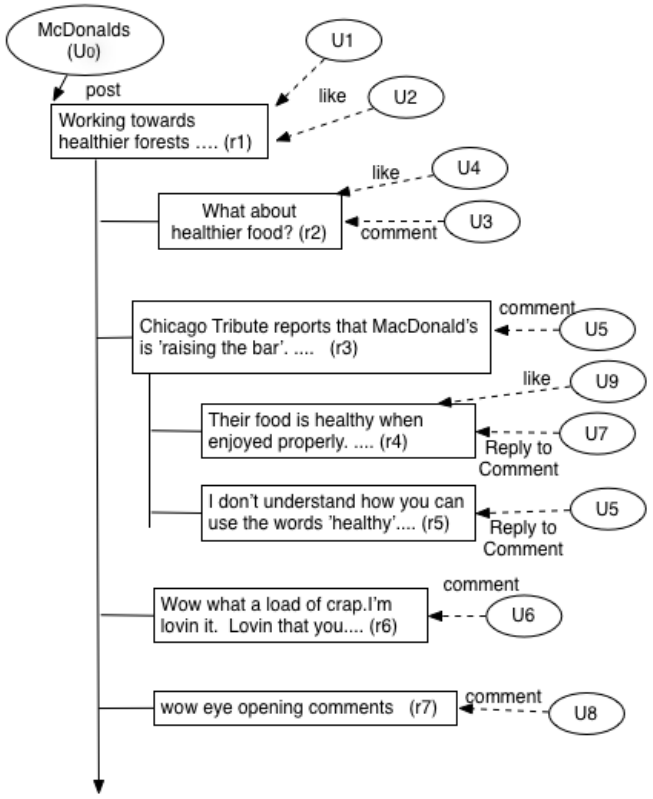


Figure 3. Facebook post example in formal model

r_1 = Working towards healthier forests through more sustainable packaging. Learn more about how McDonald's is addressing climate change: http://McD.to/6188BrQzM

r_2 = What about healthier food?

r_3 = Chicago Tribute reports that MacDonald's is 'raising the bar'. You mean bars with nails in them to beat live chickens with? MacDonald's is one big lie. Don't believe them. Next they'll tell you their food is healthy.

r_4 = Their food is healthy when enjoyed properly. Their beef is amazing and that's what they move a lot of. The fattier menu items, if you have any modicum of a pallet, you'll notice are sides and not to be enjoyed in such an amount as whole meals themselves, but hey, I know some people who think raw sugar is a treat.

r_5 = I don't understand how you can use the words 'healthy' and MacDonald's in the same sentence. They manufacture (and I use that word deliberately) to have a perfect balance of salt, sugar and fat to hook children with their 'Happy Meals'. Sorry Keith, but healthy does not contain GMO's, Factory Farmed Animals, Chicken beaks, feathers etc, wood cellulose, fat, sugar and salt.

r_6 = Wow what a load of crap. I'm lovin it. Lovin that you are losing business and closing stores. Serving gmo.poison and promoting health. I want to puke

r_7 = wow eye opening comments

The example shown in Fig. 3 can be encoded as follows, The social Data $\mathsf{D} = (\mathsf{I}, \mathsf{C})$ contains two components: $\mathsf{I} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \triangleright, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$ is the Interactions and $\mathsf{C} = (\mathrm{To}, \mathrm{Ke}, \mathrm{Pr}, \mathrm{Se}, \rightarrow_{\mathsf{topic}}, \rightarrow_{\mathsf{key}}, \rightarrow_{\mathsf{pro}}, \rightarrow_{\mathsf{sen}})$ is the Conversations.

Initailly, let us assume that the sets of activities, topics, keywords, pronouns and sentiments will have the following values.
$\mathrm{Ac} = \{\text{promotion}\}$,
$\mathrm{To} = \{\text{healthy food, sustainable packaging}\}$,
$\mathrm{Ke} = \{\text{healthy, sustainable, beef, chicken, . .}\}$
$\mathrm{Pr} = \{We, I\}, \mathrm{Se} = \{+, 0, -\}$,
$\mathsf{U} = \{u_0, u_1, ...\}$
$\mathsf{R} = \{r_1\}$
$\rightarrow_{\mathsf{act}} = \{(r_1, promotion)\}$
post action by u_0
$\mathsf{D} \bigoplus_p (u_0, r_1) = \mathsf{D}_1 = (\mathsf{I}_1, \mathsf{C})$ where
$\mathsf{I}_1 = (\mathsf{U}_1, \mathsf{R}_1, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \rhd, \rightarrow_{post\ 1}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$ with the following values
$\mathsf{U}_1 = \mathsf{U} \cup \{u_0\}$, $\mathsf{R} = \mathsf{R} \cup \{r_1\}$ and
$\rightarrow_{post_1} = \rightarrow_{post} \cup \{(u_0, \{r_1\})\}$
like action by u_2 **and** u_1
Let's imagine that the post was liked by user u_2 first and then liked by user u_1.
$\mathsf{D}_1 \bigoplus_l (u_2, r_1) \bigoplus_l (u_1, r_1) = \mathsf{D}_2 = (\mathsf{I}_2, \mathsf{C})$ where
$\mathsf{I}_2 = (\mathsf{U}_2, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \rhd, \rightarrow_{post\ 1}, \rightarrow_{share}, \rightarrow_{like\ 1}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$ with the following values
$\mathsf{U}_2 = \mathsf{U}_1 \cup \{u_2\} \cup \{u_1\}$, and
$\rightarrow_{like\ 1} = \rightarrow_{like} \cup \{(u_2, r_1), (u_1, r_1)\}$

comment action by u_5 **on the post** r_1
Let's imagine that the user u_5 posted a comment (r_3) on the Facebook post and let D_3 be the social data before the comment action.
$\mathsf{D}_3 \bigoplus_c (u_5, r_3, r_1) = \mathsf{D}_4 = (\mathsf{I}_4, \mathsf{C})$ where
$\mathsf{I}_4 = (\mathsf{U}_4, \mathsf{R}_3, \rhd_1, \mathsf{r}_{\mathsf{type}}, \mathrm{Ac}, \rightarrow_{post\ 3}, \rightarrow_{share}, \rightarrow_{like\ 1}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$ with the following values
$\mathsf{U}_3 = \mathsf{U}_3 \cup \{u_5\}$, $\mathsf{R}_3 = \mathsf{R}_2 \cup \{r_3\}$, $\rightarrow_{post\ 3} = \rightarrow_{post\ 2} \cup \{(u_5, \{r_3\})\}$ and $\rhd_1 = \rhd \cup \{(r_3, r_1)\}$.

Reply to comment by u_7 **on the comment** r_3
Let's imagine that the user u_7 posted a reply (r_4) on the comment (r_4).
$\mathsf{D}_4 \bigoplus_c (u_7, r_4, r_3) = \mathsf{D}_5 = (\mathsf{I}_5, \mathsf{C})$ where
$\mathsf{I}_5 = (\mathsf{U}_5, \mathsf{R}_4, \rhd_2, \mathsf{r}_{\mathsf{type}}, \mathrm{Ac}, \rightarrow_{post\ 4}, \rightarrow_{share}, \rightarrow_{like\ 1}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$ with the following values
$\mathsf{U}_5 = \mathsf{U}_4 \cup \{u_7\}$, $\mathsf{R}_4 = \mathsf{R}_3 \cup \{r_4\}$, $\rightarrow_{post\ 4} = \rightarrow_{post\ 3} \cup \{(u_7, \{r_4\})\}$ and $\rhd_2 = \rhd_1 \cup \{(r_4, r_3)\}$.

The rest of the operations shown in Fig. 3 can expressed similarly in the formal model.

VI. Case Study 1: Fuzzy-Set Based Sentiment Analysis

At the enterprise level, as Li and Leckenby [61] observed, technological advances such as the Internet have resulted in the vertical integration of business channel capacities such as production, distribution, transaction (e.g., Amazon and other e-commerce websites) and a horizontal integration of marketing functions such as advertising, promotions, public relations (e.g., Facebook and other social media platforms). At the agentic level of consumers, Internet and social media platforms resulted in changes not only to consumers' attitudes, perceptions and behaviours but also to the decision-making process itself in terms of the consideration set, search criteria, heuristics, and time [2], [62]. Taken together this led to the emergence of organizations that strategically utilize the online channels including social media platforms for business purposes [4]. This results in vast amounts of social data related to an enterprise's products, services, policies and processes. As such, one key application domain for sentiment analysis in enterprises is to monitor brand image, loyalty, and reputation.

Sentiment analysis can help in the understanding the user motivations for social media engagement, the different phases of consumer decision-making process and the potential business value and organizational impact of positive, negative and neutral sentiments. To illustrate this point, let us consider the following instance of socially shared consumption [63] : a positive mention about a product resulting from an automated status update of digital consumption on social media platform such as Facebook. In terms of consumer decision-making, this Facebook post can play a role in all three different orderings of the Hierarch of Effects (HoE) [64], [65] in terms of learning about the product, evaluating one's own experience of it with those of others, and engaging with the product as a brand loyalist by following that particular product related Facebook pages and posts. Similarly, the interactional dynamics of users sentiments on social media platforms might help companies better understand the sales funnel models such as AIDA (Attention, Interest, Desire and Action) [61]. Sentiments of users' posts might provide value in terms of social capital and/or signaling by turning the private individual act of consumption into a public social event and thereby signaling the user's characteristics such as taste, class, conscientiousness, and/or wealth. In other words, sociological dynamics and marketing implications similar to the conspicuous consumption [66].

A. Formal Model of Fuzzy Social Data

In this section, we will extend the formal semantics of social data presented in Sec. V with the semantics of Fuzzy sets. Regarding notations for the formal model of Fuzzy Social Data, we will follow the same notations mentioned in Sec. V.

1) Fuzzy Sets: First, we will recall necessary basic definitions of Fuzzy sets [67].

Definition 6.1: If X is a set of elements denoted by x, then a fuzzy set A over X is defined as a set of ordered pairs $A = \{(x, \mu_A(x)) \mid x \in X)\}$ where $\mu_A : X \rightarrow [0, 1]$ is the membership function.

Each member or element of a fuzzy set A is mapped to real number between 0 and 1 ([0,1]), which represents the degree of membership of an element in the fuzzy set. A membership value of 1 indicates full membership, while a value of 0 indicates no membership.

Definition 6.2: For a (finite) fuzzy set A, the *cardinality* is defined as $\mid A \mid = \sum_{x \in X} \mu_A(x)$, which is the summation of all membership values of a fuzzy set. The *relative cardinality* $\parallel A \parallel$ is defined as $\parallel A \parallel = \frac{|A|}{|X|}$, where $\mid X \mid$ is the number of elements in set X.

Definition 6.3: The support of a fuzzy set A is a crisp set of all $x \in X$ such that $\mu_A(x) > 0$. The crisp set of elements that belongs to fuzzy set A at least to a degree α is called α-level or α-cut is defined as $A_\alpha = \{x \mid x \in X \wedge \mu_A(x) \geq \alpha\}$.

Definition 6.4: The *Union* operation on two fuzzy sets $A = \{\big(x, \mu_A(x)\big) \mid x \in X)\}$ and $B = \{\big(x, \mu_B(x)\big) \mid x \in X)\}$ with membership functions μ_A and μ_B respectively is defined as a fuzzy set $\{(x, \mu_{A\cup B}(x)) \mid \mu_{A\cup B}(x) = Max(\mu_A(x), \mu_B(x))\}$.

Definition 6.5: A fuzzy relation R from a set A to B with its membership function $\mu_R : A \times B \to [0, 1]$ is defined as $R = \{\big((a, b), \mu_R(a, b)\big) \mid (a, b) \in A \times B\}$.

Similar to a fuzzy set, the membership function of a fuzzy relation indicates strength of its relationship. Moreover a fuzzy relation is nothing but a fuzzy set where the elements are ordered pairs of the relation.

2) Fuzzy Social Data: Following the definitions of Artifact Type 5.1, Actions 5.2 and Social Data 5.3 from Sec. V, we redefine fuzzy *Interactions* by redefining the activity relation ($\rightarrow_{\mathsf{act}}$) as a fuzzy relation as follows.

Definition 6.6: In Social Data $\mathsf{D} = (\mathsf{I}, \mathsf{C})$, fuzzy Interactions is defined as a tuple $\mathsf{I} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \rhd, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$ where

(i) $\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \rhd, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}$ are same as defined in 5.4,

(ii) $\rightarrow_{\mathsf{act}} = \{((r, a), \mu_{\rightarrow_{\mathsf{act}}}(r, a)) \mid r \in \mathsf{R}, a \in \mathrm{Ac}\}$ is a fuzzy relation mapping artifacts to activities with membership function $\mu_{\rightarrow_{\mathsf{act}}} : \mathsf{R} \times \mathrm{Ac} \to [0, 1]$

As shown in Def. 6.6-i, except $\rightarrow_{\mathsf{act}}$ relation, semantics of the rest of the items in fuzzy Interactions remain same as defined in the Interactions of core Social Data formal model. The $\rightarrow_{\mathsf{act}}$ is a fuzzy relation indicates a mapping between artifacts to activities (Def. 6.6-ii) with a membership function ($\mu_{\rightarrow_{\mathsf{act}}}$) indicating the strength of relationship, varies between 0 to 1. A membership value of 0 indicates complete non-existence of relationship between an artifact to an activity, where as value of 1 indicates full existence of such relationship. A value in between 0 to 1 indicates partial existence of the relationship.

Similarly, we define fuzzy Conversations of Social data as follows by redefining all relations as fuzzy relations.

Definition 6.7: In Social Data $\mathsf{D} = (\mathsf{I}, \mathsf{C})$ we define fuzzy Conversations as C= (To, Ke, Pr, Se, $\rightarrow_{\mathsf{topic}}, \rightarrow_{\mathsf{key}}, \rightarrow_{\mathsf{pro}}, \rightarrow_{\mathsf{sen}}$) where

(i) $\mathrm{To}, \mathrm{Ke}, \mathrm{Pr}, \mathrm{Se}$ are the sets of topics, keywords, pronouns and sentiments respectively as defined in 5.5

(ii) $\rightarrow_{\mathsf{topic}} = \{\big((r, to), \mu_{\rightarrow_{\mathsf{topic}}}(r, to)\big) \mid r \in \mathsf{R}, to \in \mathrm{To}\}$ is a Fuzzy relation mapping artifacts to topics with membership function $\mu_{\rightarrow_{\mathsf{topic}}} : \mathsf{R} \times \mathrm{To} \to [0, 1]$,

(iii) $\rightarrow_{\mathsf{key}} = \{\big((r, ke), \mu_{\rightarrow_{\mathsf{key}}}(r, ke)\big) \mid r \in \mathsf{R}, ke \in \mathrm{Ke}\}$ is a Fuzzy relation mapping artifacts to keywords with membership function $\mu_{\rightarrow_{\mathsf{key}}} : \mathsf{R} \times \mathrm{Ke} \to [0, 1]$,

(iv) $\rightarrow_{\mathsf{pro}} = \{\big((r, pr), \mu_{\rightarrow_{\mathsf{pro}}}(r, pr)\big) \mid r \in \mathsf{R}, pr \in \mathrm{Pr}\}$ is a Fuzzy relation mapping artifacts to pronouns with membership function $\mu_{\rightarrow_{\mathsf{pro}}} : \mathsf{R} \times \mathrm{Pr} \to [0, 1]$,

(v) $\rightarrow_{\mathsf{sen}} = \{\big((r, se), \mu_{\rightarrow_{\mathsf{sen}}}(r, se)\big) \mid r \in \mathsf{R}, se \in \mathrm{Se}\}$ is a Fuzzy relation mapping artifacts to sentiments with membership function $\mu_{\rightarrow_{\mathsf{sen}}} : \mathsf{R} \times \mathrm{Se} \to [0, 1]$.

The semantics of sets of *topics* (To), *keywords* (Ke), *pronouns* (Pr), and *sentiments* (Se) in fuzzy Conversations remain the same as in the case of Conversations (Def. 5.5) of Core formal model of Social data. Furthermore, one may note that all the relations in fuzzy Conversations ($\rightarrow_{\mathsf{topic}}$, $\rightarrow_{\mathsf{key}}$, $\rightarrow_{\mathsf{pro}}$ and $\rightarrow_{\mathsf{sen}}$) are defined as fuzzy relations with membership function varies from [0, 1], indicating the strength of relationships, where as the relations in Conversations of core formal model of social data are crisp relations.

B. Methodology

In this section, we will outline a method for calculating the sentiments of artifacts and actors based on formal model presented in previous section.

1) Sentiment Analysis: In contrast to the analytical focus on relationships in traditional social network analysis (SNA) methods, our analytical focus is on associations of actors and artefacts as sets and fuzzy sets based on certain criteria for actions, activities, sentiments, topics etc. In our associational approach, we model set and fuzzy set memberships of *Actors* performing *Actions* in *Activities* on *Artifacts*. Artifacts carry direct sentiment as they can be analysed by a sentiment engine and assigned a sentiment score and label by the sentiment engine. Individually, an action does not carry any sentiment, but it is the artifacts on which these actions are carried over, that contain sentiments. Similarly, even though actors does not carry sentiment directly, but they express their sentiments by performing actions on the artifacts, which contain the direct sentiment. Therefore, the sentiment attributed to an actor can be inferred or derived from the artifacts on which the actions are performed. Let us assume that the set of sentiments in the Conversations contain some predefined labels: *positive* (+), *neutral* (0) and *negative* (−) as indicated in $\mathrm{Se} = \{+, 0, -\}$.

2) Sentiment Analysis of Artifacts: In this sentiment analysis of artifacts, let us assume that we are confined to textual types of artifacts, i.e. $\mathsf{r}_{\mathsf{type}}(r) = (\text{post} \vee \text{comment})$. Using an automatic method (for example using a natural language processing engine) for categorising sentiment of artifacts, an artifact can be mapped to different sentiment labels with a score indicating probability of relevance between the artifact and sentiment label. Normally, these scores are expressed as either percentages or real numbers (between 0 to 1), and the sum of such scores of an artifact for multiple sentiment labels will be equal to 1.

Therefore, in this sentiment analysis, we consider the sentiment score of an artifact as it's membership value of relationship between an artifact and a sentiment label ($\rightarrow_{\mathsf{sen}}$). For example, if the sentiment of an artifact r_1 is categorised among three sentiment labels as 0.43 *positive*, 0.26 *neutral* and 0.31 *negative*, then it is encoded in the sentiment fuzzy relation ($\rightarrow_{\mathsf{sen}}$) as

$\rightarrow_{\mathsf{sen}} = \{., \big((r_1, +), 0.43\big), \big((r_1, 0), 0.26\big), \big((r_1, -), 0.31\big), .\}$.

Furthermore, we can perform an $\alpha-cut$ operation (Def. 6.3) on a Fuzzy set, to convert it to a crisp set containing set members, whose membership value is at least to the degree of of $\alpha \in [0, 1]$.

$$\mathsf{R}^{se}_{\alpha} = \{r \mid (\mu_{\rightarrow_{\mathsf{sen}}}(r, se) \geq \alpha)\}$$

Finally the crisp set R_{α}^{se} contains all the desired artifacts whose sentiment is more than certain minimum value (α). Based on the context and requirements, one could apply different $\alpha - cuts$ to the fuzzy set to $\rightarrow_{\mathsf{sen}}$, to get the crisp sets containing artifacts meeting to certain minimum sentiment score as criteria (α).

Especially, the method of application $\alpha - cuts$ is quite useful when we want to explore a phenomena which is very feebly represented in the data corpus. For example, in order to explore a weak negative sentiment in response to an event in the data corpus, one could go for a very low value of $\alpha - cut$ (e.g. $\alpha = 0.2$ or even less), to further analyse the data in a magnified view to get fine grained data visualisations. On there other hand, if some one wants to get a more abstract view on a dominantly represented sentiment values, adopting higher values of $\alpha - cut$ (e.g. $\alpha > 0.6$ or even more) will results in a view with a course grained data visualisations where only strong sentiments are represented.

Actors associated with Artifacts:: Several actors are associated with an artifact. For example actors can perform *post*, *comment share* and *like* actions on an artifact. Of course, actors can also perform *tag* action on an artifact, but we will ignore tagging operation for sentiment analysis in this paper. The set of actors that are associated with the given set of artifacts (e.g. R_{α}^{se}), can be computed as follows,

$$\begin{aligned}&\forall r \in \mathsf{R}_{\alpha}^{se}.\\ &\mathsf{U}_{\mathsf{R}_{\alpha}^{se}} = \{u \mid r \in \rightarrow_{post}(u)\} \cup\\ &\qquad \{u \mid r' \in \mathsf{R} \wedge r' \in \rightarrow_{post}(u) \wedge \rhd(r') = r\} \cup\\ &\qquad \{u \mid (u,r) \in \rightarrow_{share}\} \cup\\ &\qquad \{u \mid (u,r) \in \rightarrow_{like}\}.\end{aligned}$$

As formally expressed above, the set of actors ($\mathsf{U}_{\mathsf{R}_{\alpha}^{se}}$) associated with given set of artifacts (R_{α}^{se}) contains sets of users who posted the artifacts, who commented on the artifacts, who shared the artifacts and who liked the artifacts. One could notice that both the set of actors ($\mathsf{U}_{\mathsf{R}_{\alpha}^{se}}$) and set of artifacts ($\mathsf{R}_{\alpha}^{se}$) are crisp sets and taking the cardinality of these sets will provide us the number of members in them. One of the ways to analyse the sentiment over a time scale could be to compute these sets (R_{α} and $\mathsf{U}_{\mathsf{R}_{\alpha}^{se}}$) for each sentiment label ($\forall .se \in \{+, 0, -\}$) for given time span intervals to plot them across the time horizon.

3) Sentiment Analysis of Actors: As explained in the previous section, the sentiment attributed to an actor can be derived from the artifacts on which actions are performed by the actor. An actor can perform different actions: *post*, *comment*, *share*, *like* and *tag* on different artifacts. However *tag* action is not considered for the sentiment analysis as mentioned previously. From the formal model, for any given actor, we can compute the sets of artifacts over which the actor performed actions as mentioned previously. Building on that, for any given artifact we can also compute the sentiment scores associated with different sentiment labels from the sentiment relation ($\rightarrow_{\mathsf{sen}}$).

Therefore, the sentiment associated with an actor (u^{se}) can be defined as a tuple containing the following fuzzy sets,

$$(\rightarrow_{p}^{se}, \rightarrow_{c}^{se}, \rightarrow_{s}^{se}, \rightarrow_{l}^{se})$$

1) $\rightarrow_{p}^{se} = \{\big((r, se), \mu_p(r, se)\big) \mid r \in \rightarrow_{post}(u) \wedge \rhd(r)$ is not defined$\}$ is a fuzzy set containing all the artifacts that are posted by the user with $\mu_p(r, se) = \mu_{\rightarrow_{\mathsf{sen}}}(r, se)$ as membership function,
2) $\rightarrow_{c}^{se} = \{\big((r, se), \mu_c(r, se)\big) \mid r \in \rightarrow_{post}(u) \wedge \exists r' \in \mathsf{R}. \rhd(r) = r'\}$ is a fuzzy set containing all the comment artifacts that are posted by the user, with $\mu_c(r, se) = \mu_{\rightarrow_{\mathsf{sen}}}(r, se)$ as membership function,
3) $\rightarrow_{s}^{se} = \{\big((r, se), \mu_s(r, se)\big) \mid (u, r) \in \rightarrow_{share}\}$ is a fuzzy set containing all the artifacts that are shared by the user, with $\mu_s(r, se) = \mu_{\rightarrow_{\mathsf{sen}}}(r, se)$ as membership function,
4) $\rightarrow_{l}^{se} = \{\big((r, se), \mu_l(r, se)\big) \mid (u, r) \in \rightarrow_{like}\}$ is a fuzzy set containing all the artifacts that are liked by the user, with $\mu_l(r, se) = \mu_{\rightarrow_{\mathsf{sen}}}(r, se)$ as membership function,

where $r \in \mathsf{R}$, $se \in \mathrm{Se}$, $\mu_{\rightarrow_{\mathsf{sen}}}(r, se)$ is the membership function of the sentiment fuzzy relation ($\rightarrow_{\mathsf{sen}}$).

The the sentiment associated with an actor (u^{se}) can calculated by application of union operation (Def. 6.4) on the above fuzzy sets ($\rightarrow_{p}^{se} \cup \rightarrow_{c}^{se} \cup \rightarrow_{s}^{se} \cup \rightarrow_{l}^{se} \cup \rightarrow_{t}^{se}$). Therefore, sentiment associated with an actor (u^{se}) can be computed as follows

$$u^{se} = \{\big((r, se), \mu_u(r, se)\big) \mid r \in \mathsf{R}_u\}, \text{ where}$$

1) R_u is set of artifacts for an actor (u) over which the actions are performed
   $\mathsf{R}_u = \rightarrow_{post}(u) \cup \rightarrow_{share}(u) \cup \rightarrow_{like}(u)$.
   Notice that, the set $\rightarrow_{post}(u)$ contains all artifacts that are posted and commented by the user.
2) the membership function is defined as
   $\mu_u(r, se) = \mathrm{Max}(\mu_p(r, se), \mu_c(r, se), \mu_s(r, se), \mu_l(r, se))$

One could observe that the associated sentiment of an actor is a fuzzy set with artifacts and sentiment labels with membership values as the sentiment scores. Therefore, one could apply the $\alpha - cuts$ on the fuzzy set to extract a crisp set (u_{α}^{se}) meeting up the criteria for each sentiment label ($\forall .se \in \{+, 0, -\}$).

Furthermore, the same method can applied to get such sets for different time span intervals with in a time period. One of the ways to analyse the associated actor sentiment over a time scale could be to compute these sets (u_{α}^{se}) for each sentiment label ($\forall .se \in \{+, 0, -\}$) for given time span intervals and plot their cardinalities (e.g. number of artifacts in the set for $+$ sentiment) across the time horizon. In this way, we could profile the associated sentiment of an actor over a period of time by computing how the cardinalities of the sets of the associated sentiment labels of an actor varies over timeline.

C. Illustrative Example

In this section, we will exemplify the formal model with fuzzy sets by taking an example post from the Facebook page of H&M cloth stores as shown in the figure 4. In order to enhance the readability of the example, the artifacts (e.g. texts) have been annotated as $r1, r2$ etc and the annotated values will be used in encoding the example using the formal model.

Moreover, as our focus is to mainly to demonstrate sentiment analysis, we will abstract away from the details of the sets (e.g. Topics, Keywords etc) which are not directly involved in the sentiment analysis. As shown in Figure 4,

the sentiments of the artifacts (e.g. (+):20, (0):65, (-):15) are represented in the boxes below the artifacts.

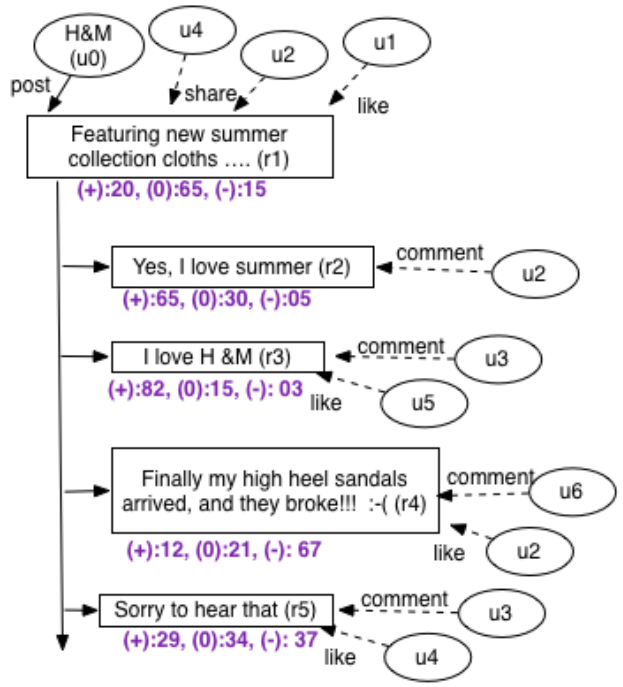


Figure 4. Example in formal model

Example 6.1: The example shown in Fig. 4 will be encoded as follows,
$\mathsf{D} = (\mathsf{I}, \mathsf{C})$ where $\mathsf{I} = (\mathsf{U}, \mathsf{R}, \mathrm{Ac}, \mathsf{r}_{\mathsf{type}}, \rhd, \rightarrow_{post}, \rightarrow_{share}, \rightarrow_{like}, \rightarrow_{tag}, \rightarrow_{\mathsf{act}})$ is the Interactions and C= (To, Ke, Pr, Se, $\rightarrow_{\mathsf{topic}}, \rightarrow_{\mathsf{key}}, \rightarrow_{\mathsf{pro}}, \rightarrow_{\mathsf{sen}}$) is the Conversations.

Initially, the sets of actors, artifacts and other relations have the following values.
$\mathsf{U} = \{u_0, u_1, u_2, u_3, u_4, u_5, u_6, ...\}$
$\mathsf{R} = \{r_1, r_2, r_3, r_4, r_5, ...\}$
$\rhd = \{(r_2, r_1), (r_3, r_1), (r_4, r_1), (r_5, r_1), ..\}$
$\rightarrow_{post} = \{(u_0, \{r_1, ..\}), (u_2, \{r_2\}), (u_3, \{r_3, r_5\}), (u_6, \{r_4\}), ..\}$
$\rightarrow_{share} = \{(u_4, r_1), (u_2, r_1)..\}$
$\rightarrow_{like} = \{(u_1, r_1), (u_5, r_3), (u_2, r_4), (u_4, r_5), ..\}$
$\mathrm{Se} = \{+, 0, -\}$
After the artifacts are analysed for the sentiments, the sentiment relation becomes a fuzzy set contain the pairs of artifacts and sentiment labels with the sentiment score as membership value as shown below,

$$\rightarrow_{\mathsf{sen}} = \{ \big((r_1, +), 0.20\big), \big((r_1, 0), 0.65\big), \big((r_1, -), 0.15\big), \big((r_2, +), 0.65\big), \big((r_2, 0), 0.30\big), \big((r_2, -), 0.05\big), \big((r_3, +), 0.82\big), \big((r_3, 0), 0.15\big), \big((r_3, -), 0.03\big), \big((r_4, +), 0.12\big), \big((r_4, 0), 0.21\big), \big((r_4, -), 0.67\big), \big((r_5, +), 0.29\big), \big((r_5, 0), 0.34\big), \big((r_5, -), 0.37\big)\}$$

Regarding temporal dimension ($\mathbf{T}$), let us assume that the post (in Figure 4) and all its conversation happened in same time frame ($t_1 - t_2$), then sentiment relation for time period ($t_1 - t_2$) is same as $\rightarrow_{\mathsf{sen}}$.

From the sentiment fuzzy set, one can extract different crisp sets (R^{se}_{α}) for artifacts based different values of $\alpha - cuts$. For example for a value of $\alpha = 0.4$, the artifact sets for $+$ and $-$ will be
$\mathsf{R}^{+}_{\alpha=0.40} = \{r_2, r_3\}$ and $|\mathsf{R}^{+}_{\alpha=0.40}| = 2$
$\mathsf{R}^{-}_{\alpha=0.40} = \{r_4\}$ and $|\mathsf{R}^{-}_{\alpha=0.40}| = 1$

On the other hand, if some one wants a fine grained analysis of the data, they could use a lower value for $\alpha - cut$, which will include more elements into the analysis.
$\mathsf{R}^{+}_{\alpha=0.20} = \{r_1, r_2, r_3, r_5\}$ and $|\mathsf{R}^{+}_{\alpha=0.20}| = 4$
$\mathsf{R}^{-}_{\alpha=0.20} = \{r_4, r_5\}$ and $|\mathsf{R}^{-}_{\alpha=0.20}| = 2.$
Similarly, we can also compute the actor sets ($\mathsf{U}_{\mathsf{R}^{se}_{\alpha}}$) that are associated with the artifact sets as follows.

$$\begin{aligned}\mathsf{U}_{\mathsf{R}^{+}_{\alpha=0.40}} &= \{u_2\} \cup \emptyset \cup \emptyset \cup \emptyset \cup \{u_3\} \cup \emptyset \cup \emptyset \cup \{u_5\} \\ &= \{u_2, u_3, u_5\} \\ \mathsf{U}_{\mathsf{R}^{-}_{\alpha=0.20}} &= \{u_6\} \cup \emptyset \cup \emptyset \cup \{u_2\} \cup \{u_3\} \cup \emptyset \cup \emptyset \cup \{u_4\} \\ &= \{u_6, u_2, u_3, u_4\}\end{aligned}$$

Notice that, here we have an advantage due to fuzzy set modelling that an can an actor can be present in more than one set (e.g. $\mathsf{U}_{\mathsf{R}^{+}_{\alpha=0.2}}$ and $\mathsf{U}_{\mathsf{R}^{-}_{\alpha=0.2}}$), as an actor can express more than one sentiment by performing the actions on artifacts in reality. When once crisp sets for artifacts (R^{se}_{α}) and actors ($\mathsf{U}_{\mathsf{R}^{se}_{\alpha}}$) are computed on a time scale for given time spans, one can plot their cardinalities against the time scale.

1) Inferred Sentiment and Actor Profiling:: As explained in the previous section, the inferred sentiment for actors can be calculated in the similar line as above. In this example, we will show how one can compute inferred sentiment for the actor u_2, where we take union of fuzzy sets containing artifacts with sentiment labels for the artifacts posted, shared and liked by actor u_2 as follows.

$$\begin{aligned}u_2^{+} &= \{\big((r_2, +), 0.65\big)\} \cup \{\big((r_1, +), 0.20\big)\} \cup \{\big((r_4, +), 0.12\big)\} \\ &= \{\big((r_2, +), 0.65\big), \big((r_1, +), 0.20\big), \big((r_4, +), 0.12\big)\} \\ u_2^{-} &= \{\big((r_2, -), 0.05\big)\} \cup \{\big((r_1, -), 0.15\big)\} \cup \{\big((r_4, -), 0.67\big)\} \\ &= \{\big((r_2, -), 0.05\big), \big((r_1, -), 0.15\big), \big((r_4, -), 0.67\big)\}\end{aligned}$$

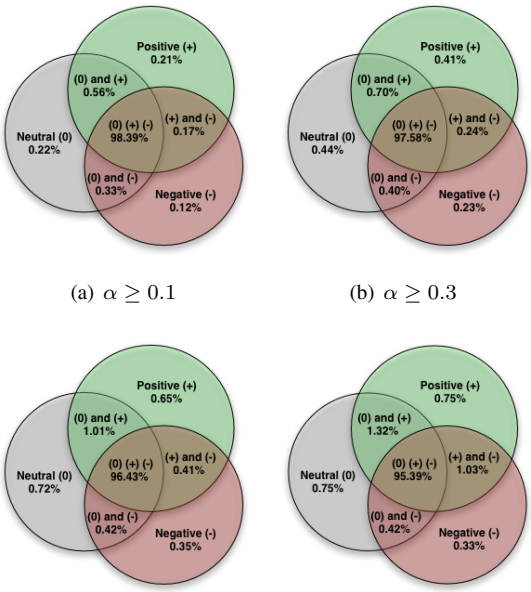
After computing the fuzzy sets as above, one could apply $\alpha - cut$ with the required granularity to get crisp sets similar to the sentiment analysis of the artifacts. After that many such sets can be computed for a given time intervals and can be plotted on a time scale to analyse how the sentiment of an actor varies in the time frame.

D. Case Study and Findings

In this section, we present a case study where big social data of the fast fashion company, H&M is collected from its Facebook page. We empirically analyse the sentiment of artifacts on social data collected by Social Data Analytics Tool (SODATO) [38] from the Facebook page of H&M and analysis using the methodology presented in the previous section that is based on formal modelling of social data.

1) Conversation Analysis: Google Prediction API [68] was utilized in order to calculate sentiments for the posts and comments on the wall. Google Prediction API provides RESTful API access to the service. Configuration for computation of sentiment began with the setting up a model which was trained with the manually labelled data subset from the H&M data corpus fetched by SODATO. This training dataset consisted of 11,384 individual posts and comments randomly selected from H&M data corpus and their corresponding sentiment

labels as coded by five different student analysts. Training data was labelled *Positive*, *Negative* or *Neutral* and the file was uploaded on the Google Cloud Storage using the console explorer interface provided by the Google.



(a) $\alpha \geq 0.1$ (b) $\alpha \geq 0.3$

(c) $\alpha \geq 0.5$ (d) $\alpha \geq 0.7$

Figure 5. Artifact sentiments

After successful training of the model, Sentiment module provided by SODATO was utilized to calculate sentiment for posts and comments for the entire conversations corpus of H&M. The sentiment results for each individual post/comment returned by the Google Prediction API were saved back to the relational database. In order to calculate quarterly aggregation of the sentiment classified conversations, further segmentation and grouping was performed using SQL queries and relational database entities were used to store data and it was made available for Analytical calculations.

	α-cuts				
sentiment	$\geq$ **0.1**	$\geq$ **0.3**	$\geq$ **0.5**	$\geq$ **0.7**	$\geq$ **0.9**
$+$	17,752	25,949	30,343	25,869	19,974
$-$	9,166	14,503	16,577	13,494	10,397
0	12,566	21,607	26,826	24,312	21,830
$+ \cap -$	5,661	5,184	2,067	1,489	913
$+ \cap 0$	16,674	14,401	8,550	7,069	4,673
$- \cap 0$	10,017	9,984	6,541	5,381	3,892
$+ \cap - \cap 0$	39,001	19,209	6,512	4,567	2,739
Total artifacts	110,837	110,837	97,416	82,181	64,418

Table III
PARENT ARTIFACT (POSTS) SENTIMENT DISTRIBUTION

2) Data Analysis: The H&M Facebook wall was fetched for a time period from 12-March-2007 to 31-December-2013 using SODATO tool. The total data corpus for that period contains 12.58 million data elements including posts, comments, likes on posts and comments and shares. The sentiment scores for the 12.58 million data elements were analysed using Google Prediction API [68].

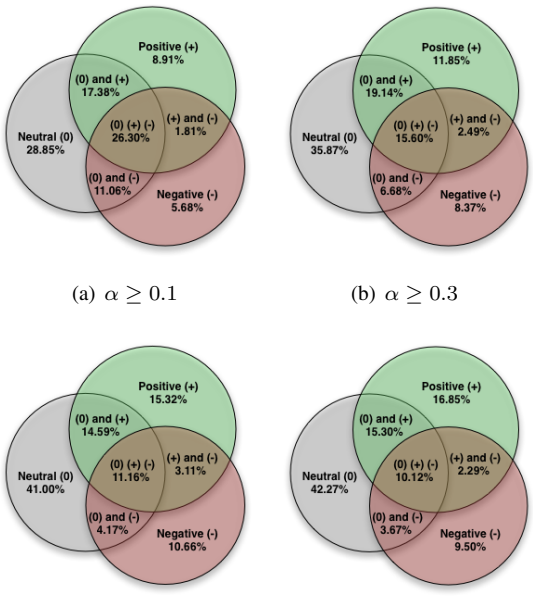
	α-cuts				
sentiment	$\geq$ **0.1**	$\geq$ **0.3**	$\geq$ **0.5**	$\geq$ **0.7**	$\geq$ **0.9**
$+$	36,114	57,653	77,551	80,540	83,378
$-$	19,433	32,472	42,145	35,388	28,310
0	37,511	62,037	85,334	80,404	79,981
$+ \cap -$	28,788	33,929	49,315	109,785	237,156
$+ \cap 0$	94,094	99,339	119,822	141,158	297,516
$- \cap 0$	54,756	56,527	50,176	44,660	35,520
$+ \cap - \cap 0$	16,537,774	13,810,588	11,477,670	10,189,815	7,742,858
Total artifacts	16,808,470	14,152,545	11,902,013	10,681,750	8,504,719

Table IV
TOTAL ARTIFACT (POSTS + COMMENTS + LIKES) SENTIMENT DISTRIBUTION

	α-cuts				
sentiment	$\geq$ **0.1**	$\geq$ **0.3**	$\geq$ **0.5**	$\geq$ **0.7**	$\geq$ **0.9**
$+$	331,891	441,290	549,159	563,600	555,964
$-$	211,783	311,861	382,082	317,912	199,815
0	1,074,602	1,335,933	1,469,989	1,413,921	1,168,176
$+ \cap -$	67,496	92,901	111,438	76,491	51,868
$+ \cap 0$	647,315	712,828	523,046	511,667	508,537
$- \cap 0$	411,821	248,707	149,532	122,645	66,889
$+ \cap - \cap 0$	979,718	581,106	400,186	338,565	231,158
Total actors	3,724,626	3,724,626	3,585,432	3,344,801	2,782,407

Table V
ACTORS SENTIMENT WITH DIFFERENT $\alpha - cuts$

3) Findings: Compared to existing sentiment analysis methods and tools in academia and industry, the set theory and fuzzy set theory approach that we demonstrated in the tables (III, IV and V) and figures (5, 6 and 7) above reveal the longitudinal sentiment profiles of actors and artefacts for the entire corpus. The α-cut approach to sentiment analysis allows analysts (marketing professionals and/or academic researchers) to specify their own probability level for sentiment

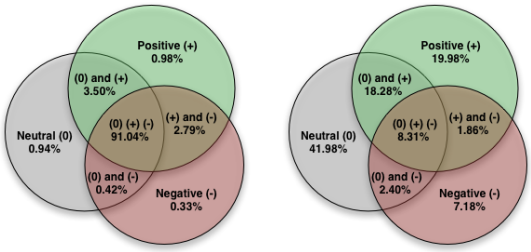


(a) $\alpha \geq 0.1$ (b) $\alpha \geq 0.3$

(c) $\alpha \geq 0.5$ (d) $\alpha \geq 0.7$

Figure 6. Actor sentiments

categories of positive, negative and neutral. Further, it allows the individual analyst to identify the intersections of positive, negative, and neutral sentiment for any given α-cut. This allows the analyst to identify strong-weak expressions of positive, negative, and neutral sentiment.



(a) Artifact sentiments $\alpha \geq 0.9$ (b) Actor sentiments $\alpha \geq 0.9$

Figure 7. Artifact and Actor sentiments for $\alpha \geq 0.9$

For example, let us consider the α-cut of 0.9 for actors in Table V and Figure 7(b). The graph shows that 7.18% of the entire Facebook user group for the company are always expressing negative sentiments whereas 19.9% of the user group is always expressing positive sentiments. With the caveat that not all of those positive and negative sentiments could be about the company itself (they could be directed towards other brands and/or other social actors on the Facebook page of the company), the results can help identify the strong brand loyalists (always positive) and strong brand critics (always negative). Similar analysis for the α-cut of 0.1 for actors will yield weak brand loyalists and critics.

With respect to the sentiment analysis of artifacts, at the α-cut of 0.9 (Table III and Figure 7(a)), we find that 0.33% of all conversations on the Facebook page were entirely negative. A quick test for the social media marketing effectiveness can be constructed by extracting the number of completely negative conversations started by the company itself. That is, it is marketing problem if the company's posts are being categorized as negative sentiment and the all ensuing interactions by its Facebook users are also negative. This might have implications for the brand reputation and image even discounting attempts at humor by self-depreciation and/or irony. Applying the crisp set and fuzzy set modelling of sentiments of actors and artefacts over critical time periods can reveal the temporal dynamics of how different users express their sentiments for different products, campaigns, and events. Having said that, our primary objective in this paper is not to provide a detailed interpretation of the results but to propose and demonstrate a new approach to sentiment analysis.

VII. Case Study 2: Social Set Analysis of Corporate Social Media Crises on Facebook

Corporate crises are by nature unpredictable but post hoc, crises appear not unexpected. Corporate crisis can trigger negative reactions from stakeholders and thereby affect the overall performance of the company. Therefore, it is important for the companies to respond to the crises in order to limit the damage [69], [70]. This paper addresses the topic of corporate crises on social media channels. Social media crises pose significant challenges for organizations in terms of their rapid propagation and deterioration of brand parameters that can have sustained negative business impacts. This paper addresses the following research questions.

Research Questions:

1) What were the characteristics of big social data before, during, and after the corporate crisis?
2) What strategies and tactics does a companies employ, if at all they do, in order to manage the social media crisis?
3) How can a company in general best manage a social media crisis?

A. Selected Corporate Social Media Crises

In order to address the above research questions, we selected four recent social media crises. The objective was to uncover temporal dyanamics and interactional patterns of big social data and to investigate the strategies and tactics adopted by the companies that have experienced social media crisis in order to manage them. We purposefully limited the selection of social media crises to Denmark and the social media platform to Facebook to hold invariant the technological,linguistic and socio-cultural aspects of interacting with social media [32], [37] invariant : Copenhagen Zoo, Telenor, Jensen's Bøfhus (translation: Jensens Steak House), and Imerco. Next, we briefly describe each corporate social media crisis.

Copenhagen Zoo experienced a social media crisis, which started on February 8th 2014, due to an impending euthanizing of a young giraffe they had chosen to call Marius and lasted until February 13th 2014. Also, major international media has also participated in the case of Marius. British BBC and The Guardian newspaper has also referred to the killing, CNN followed the case on both network and TV, and The New York Times has also written about Marius' death [71].

Telenor experienced a social media crisis on Facebook, which started on August 3rd 2014 and lasted until August 8, 2014, due to a farewell salute from an unsatisfied customer who wrote in the evening on August 2nd 2014 at Telenor's Facebook page that he had ended his mobile subscription with the telecom company. In his post, the dissatisfied customer described that Telenor could not manage to collect money by Direct Debit and that the company had repeatedly sent reminders before he had received the normal expense. This post brought Telenor into a social media crisis on Facebook [1] and more than 30,000 "liked it" [2].

Jensen's Bøfhus experienced a social media crisis on Facebook, which started on September 19, 2014 and lasted until September 27, 2014, due to a dispute between Jensen's Bøfhus, and a fish restaurant named *Jensens Fiskerestaurant* (ed. Jensen's Seafood Restaurant). The case involved a conviction in the Supreme Court that caused great debate in Denmark, since Jensen's Bøfhus were successful at that the name, Jensen Fiskerestaurant, is too similar to the steakhouse chain restaurant. This meant that the owner of Jensen's Fiskerestaurant, Jacob Jensen, had to change the name of his restaurant.

[1] Telenor on tv2.dk
[2] Telenor on politiken.dk

According to Jensen's Bøfhus they were trying to protect their trademark in the catering industry as Jensens Fiskerestaurant were planning to expand with new restaurants in other cities [3]. According to the judgment, the small restaurateur, Jacob Jensen, had to pay 200,000 Danish kroner to Jensen's Bøfhus, 150,000 Danish kroner to the costs that Jensen's Bøfhus have had his own lawyer and other expenses[4].

Imerco experienced a social media crisis, which started on August 25th, 2014 and lasted until August 26th 2014, due to a fast sold out anniversary vase from the brand Kähler. 16,000 customers wanted to buy a special anniversary vase from the company Kähler on offer at Imerco's website. However, this tumbled the website, after which angry customers vented their displeasure on Imerco's Facebook page [5].

B. Methodology

In this section, we will outline the methodology adopted to conduct big social data analytics on the Facebook walls of the companies based on the formal definitions of social data as defined in Sec. V. In the analysis, we also distinguish between admin-actor (denoted by u_a), who manages the Facebook wall of an enterprise from non-admin actors (denoted by $u \in \mathsf{U} \setminus u_a$), who are the social media users. To simply the matters, we have excluded *share* action from our analysis as we did not noticed any share actions in the datasets. Moreover, the terms *user* and *actor* are used interchangeably throughout the paper without any difference in semantics.

1) Artifact Analysis (Crisis Detection): Social media crises are characterized by marked increase in interaction levels on the social media channels. Further, based on traditional crisis communication and management theories and frameworks discussed earlier, we conducted temporal analysis of interactions in terms of two kinds of actions (like and comment) with respect to two kinds of artifacts (posts and comments) made by two different kinds of actors (admins/companies and non-admins) over temporal dimension of daily, weekly and yearly as further explained below.

artifact (post) by	actions	by actor
admin actor	comment or like	admin actor
non-admin actor		non-admin actor

Table VI
ACTIONS OF ADMIN/NON-ADMIN ACTORS ON POST ARTIFACT

Post artifact Analysis: As shown in Table VI, kind of actions that can be performed on a post artifact are comment and like. As one of the possible interactions in Table VI, comment and like actions made by non-admin actors over the post artifact created by the admin-actor can be defined as,

1) Comments by non-admin actors on admin-actor posts: $\mathsf{R}_c^{u|u_a} = \{r_c \mid (u_a, r_p) \wedge (u, r_c) \in \rightarrow_{post}\}$

2) Likes by non-admin actors on admin-actor posts: $\mathsf{L}^{u|u_a} = \{(u, r_p) \mid (u_a, r_p) \in \rightarrow_{post} \wedge (u, r_p) \in \rightarrow_{like}\}$.

[3] Jensen's Bøfhus on tv2.dk
[4] Jensen's Bøfhus on politiken.dk
[5] Imerco on politiken.dk

The set ($\mathsf{R}_c^{u|u_a}$) contains comment artifacts (r_c) made by non-admin actors (u) on the post artifact (r_p) created by admin-actor (u_a). Similarly, the set $\mathsf{L}^{u|u_a}$ contains pairs of non-admin actors (u) with their liked post artifacts (r_p), that were created by the admin-actor (u_a). Finally, total number of actions made by the non-admin actors on admin-actor posts can be calculated by taking sum of set cardinalities ($|\mathsf{R}_c^{u|u_a}| + |\mathsf{L}^{u|u_a}|$). Using this method, we have calculated weekly distribution of actions made by non-admin actors over the admin posts for the case study companies. As an example, such a distribution for Copenhagen Zoo crisis is plotted as shown in Figure 8(a). The other interactions from Table VI can be defined similarly.

Comment Artifact Analysis: **Like** is the only type of interaction that can be performed on a comment artifact. Therefore, we have conducted temporal analysis of like action (by admin vs non-admin actors) on comments made (by admin vs non-admin actors) over the posts (made by admin vs non-admin actors) on a temporal dimension of daily, weekly and yearly as are shown in Table VII.

artifact (post) by	artifact (comment) by	action
admin actor	admin	like by admin/non-admin actor
	non-admin	
non-admin actor	admin	
	non-admin	

Table VII
LIKES ON COMMENTS BY ADMIN/NON-ADMIN ACTORS OVER POSTS BY ADMIN/NON-ADMIN ACTORS

As one of the possible interactions from Table VII, we define likes by non-admin actors on comments made by non-admin actors over posts by admin-actor as follows.
Let $u_1, u_2 \in \mathsf{U} \setminus u_a$ be the non-admin actors, $r_p, r_c \in \mathsf{R}$ be the post and comment artifacts such that the comment is made on post ($r_p \rhd r_c$), then
$\mathsf{L}^{u|u|u_a} = \{(u_2, r_c) \in \rightarrow_{like} \mid (u_a, r_p), (u_1, r_c) \in \rightarrow_{post}\}$.
The set $\mathsf{L}^{u|u|u_a}$ indicates likes by non-admin actors (u) on the comments (r_c) made by non-admin actors (u) over the posts (r_p) made by admin actor (u_a).
Similarly, the likes by non-admin actors on the comments made by the admin-actor over the admin posts can be defined as, $\mathsf{L}^{u|u_a|u_a} = \{(u_1, r_c) \in \rightarrow_{like} \mid (u_a, r_p), (u_a, r_c) \in \rightarrow_{post}\}$.
Using the above methodology, comparison of likes on comments made by admin actor verses non-admin actors over the admin posts for the Jensen Bøfhus company is computed and plotted as shown in Figure 8(c).

2) Actor Analysis (Social Set Analysis): As part of social set analysis, sets containing unique actors who performed interactions *during* (U_d), *before* (U_b) and *after* (U_a) the crisis period are computed. Let $\mathbf{ts}_d$, $\mathbf{ts}_b$ and $\mathbf{ts}_a$ be time spans for *during*, *before* and *after* the crisis respectively containing respective sets of time stamps for those periods. In the social set analysis conducted on the four companies presented in this paper, we observed that the crisis period spans around two weeks on the social media platforms, therefore timespan $\mathbf{ts}_d$ contains time stamps belonging two weeks of the crisis period, where as $\mathbf{ts}_b$, $\mathbf{ts}_a$ contains timestamps belonging to two weeks before the start of the crisis and two weeks after the end of the crisis respectively.

Actors analysis for crisis period: The *during* (U_d) actors set contains the actors who have either posted or commented or liked an artifact during the crisis period ($\mathbf{ts}_d$), as defined below. Let $ac \in \{post, comment\}$, then

$$\mathsf{U}_d = \{u \mid \exists r \in \mathsf{R}.(u,r) \in \rightarrow_{post} \ \wedge \mathbf{T}(u,r,ac) \in \mathbf{ts}_d\} \ \cup \\ \{u \mid \exists r \in \mathsf{R}.(u,r) \in \rightarrow_{like} \ \wedge \mathbf{T}(u,r,like) \in \mathbf{ts}_d\}$$

where $\mathbf{T}(u,r,ac)$ and $\mathbf{T}(u,r,like)$ are timestamps of the respective actions. As indicated above the set U_d contains all the unique actors that have performed either a *post*, or a *comment* or a *like* on an artifact during the crisis period. Similarly, the unique actor sets U_b and U_a can computed where the time stamp of the actions belongs to time spans: before ($\mathbf{ts}_b$) and after ($\mathbf{ts}_a$) the crisis period respectively. Finally intersections between actor sets ($\mathsf{U}_d, \mathsf{U}_b, \mathsf{U}_a$) have been computed to represent actor Venn diagrams as shown in Fig. 10. As an example, the set of unique actors who have performed actions only during crisis (neither before nor after) can be computed using the principle of Venn diagram as: $\mathsf{U}_d \cup (\mathsf{U}_d \cap \mathsf{U}_b \cap \mathsf{U}_a) \setminus \big((\mathsf{U}_d \cap \mathsf{U}_b) \cup (\mathsf{U}_d \cap \mathsf{U}_a)\big)$.

3) Actor Analysis for likes on admin posts: The *like* action on a post is an indication of definitive support over the opinion expressed by the post. The sets of unique actors who performed like actions on the posts made by admin actor (u_a) during the crisis period (U^l_d) is computed as follows.

$$\mathsf{U}^l_d = \{u \mid \exists r \in \mathsf{R}.(u_a,r) \in \rightarrow_{post} \ \wedge (u,r) \in \rightarrow_{like} \\ and \ \mathbf{T}(u,r,like), \mathbf{T}(u_a,r,post) \in \mathbf{ts}_d\}$$

As defined above the set U^l_d contains the unique actors who have performed like action on the posts made by the admin actor on the Facebook wall of the enterprise. In the similar lines, the set of unique actors who liked the admin posts before (U^l_b) and after (U^l_a) the crisis period can be computed by considering the timestamps belonging to $\mathbf{ts}_b$ and $\mathbf{ts}_a$ time periods respectively. The Venn diagrams representing the sets of unique actors who liked admin posts are computed for four companies and shown in Fig. 11.

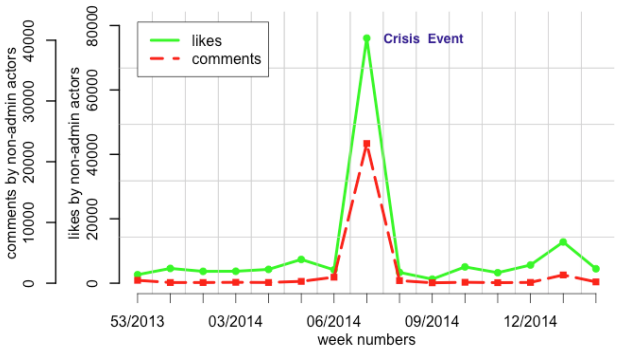
4) Actor Analysis for comments on admin posts: Unlike the *like*, the *comment* action is not a definitive action in support of the opinion expressed by a post. Therefore, we have computed the sets of unique actors who commented on the posts made by admin actor (u_a) during the crisis period (U^c_d) as follows.

$$\mathsf{U}^c_d = \{u \mid \exists r_p, r_c \in \mathsf{R}.r_p \rhd r_c \wedge (u_a,r_p), (u,r_c) \in \rightarrow_{post} \\ \wedge \ \mathbf{T}(u_a,r_p,post), \mathbf{T}(u,r_c,comment) \in \mathbf{ts}_d\}$$

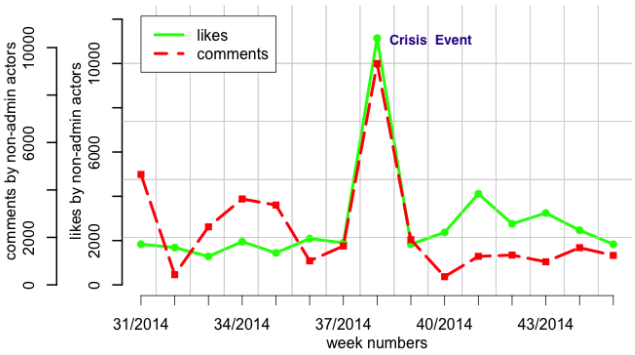
As shown above, the set U^c_d contains all the unique actors who have commented on the posts made by the admin actor (u_a) during the crisis period. The other two sets: U^c_b, U^c_a containing the unique actors who have commented on the posts made by admin actor (u_a) before and after the crisis can be computed in the similar lines. The Venn diagrams containing the intersections of the actors who have commented on admin posts before, during and after the crisis period can be computed as shown in Fig. 12.

C. Findings

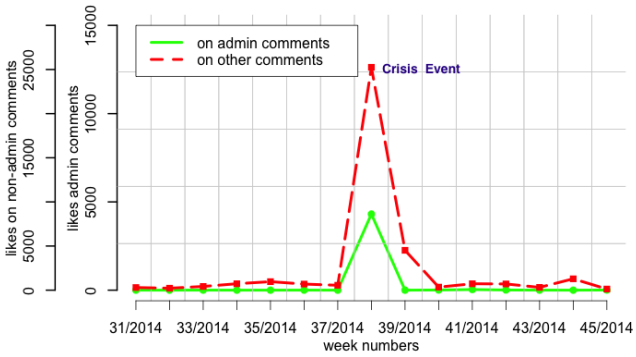
In this section, we first present the interactional patterns revealed by the Social Set Analysis and deeper substantive analysis of the social data using netnograpic analysis, manual sentiment analysis and topic discovery.



(a) Zoo - comments, likes by non-admin actors on admin posts



(b) Jensen Bøfhus - comments, likes by non-admin actors on admin posts

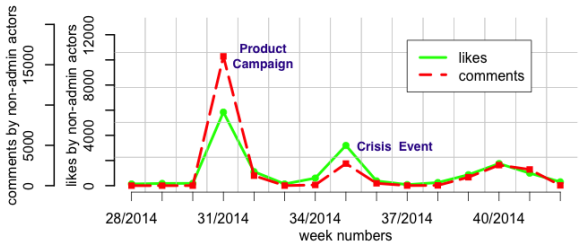


(c) Jensen Bøfhus - comparison of likes on comments made by admin vs non-admin actors

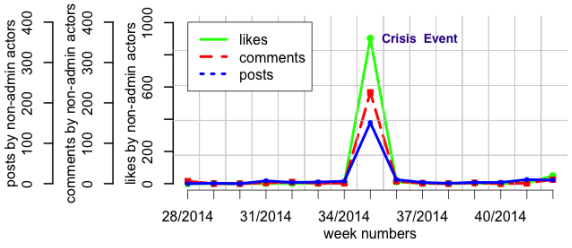
Figure 8. Artifact Analysis of Copenhagen Zoo and Jensen Bøfhus Crises [72]

1) Crisis Detection: Figures 8 and 9 present the results from the temporal analysis of interactions. Figure 8(a) reveals the interactional spikes by non-admin actors on the Copenhagen Zoo's posts as well as an preliminary indication of the nature of the crises. To be specific, the spike of likes on the admin's posts and comments is an indicator of positive endorsement of the Copenhagen Zoo's activities during the crises. As can be in seen from figure 8(c), in case of Jensen's Bøfhus, the admin comments received far less number of likes when compared to likes on comments made by non-admin users during the interactional peak, which is an indicator of negative endorsement of Jensen's Bøfhus activities.

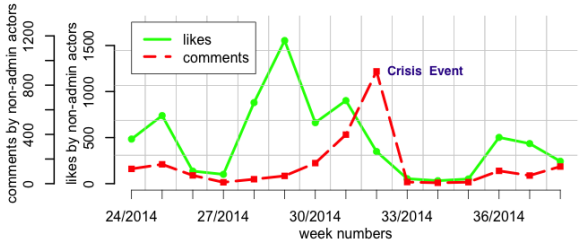
Thus, we can not only detect the interactional peaks (in this case, known social media crises) but also obtain preliminary



(a) Imerco - comments, likes by non-admin actors on admin posts



(b) Imerco - posts, comments, likes by non-admin actors on non-admin posts



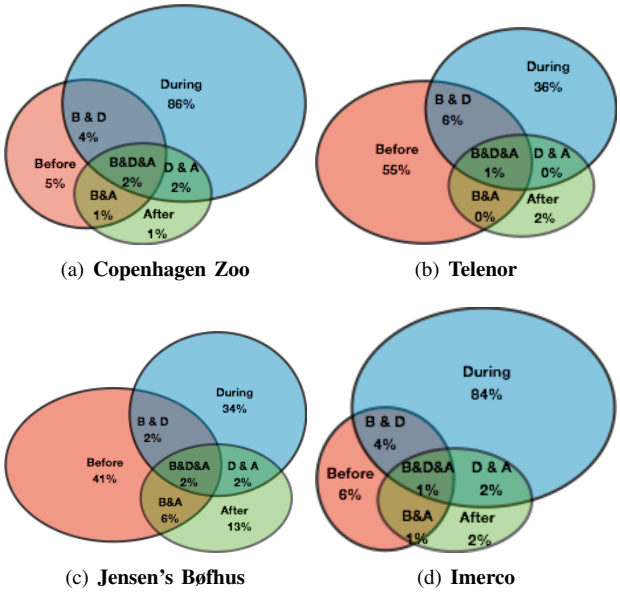
(c) Telenor - comments, likes made by non-admin actors on admin posts

Figure 9. Artifact Analysis of Imerco and telenor Crises [72]

indicators of the nature of net user sentiments towards the companies duing the crises. We supplement this with a deeper analysis of users' actions before, during and after the crises, a netnographic analysis of the Facebook walls, and sentiment and topic analysis of the posts and comments during the crises as presented and discussed next.

2) Social Set Analysis: The analytical objective for conducting Social Set Analysis (SSA) was to identify the structural properties of social media crises with reference to the domain-specific theories of crisis communication and management discusses in the theoretical framework section. Specifically, we were interested in the three time-periods of before, during and after crisis. We conducted SSA across the three time-periods for (a) overall distribution of user actions (Figure 10), (b) distribution of likes by facebook users on the artefacts (posts and comments) created by the company (Figure 11), and (c) distribution of comments by facebook users on posts created by the company (Figure 12).

As can be seen in Figure 10, a disproportionately high



(a) **Copenhagen Zoo** (b) **Telenor**

(c) **Jensen's Bøfhus** (d) **Imerco**

Figure 10. Social Set Analysis of actors during crisis

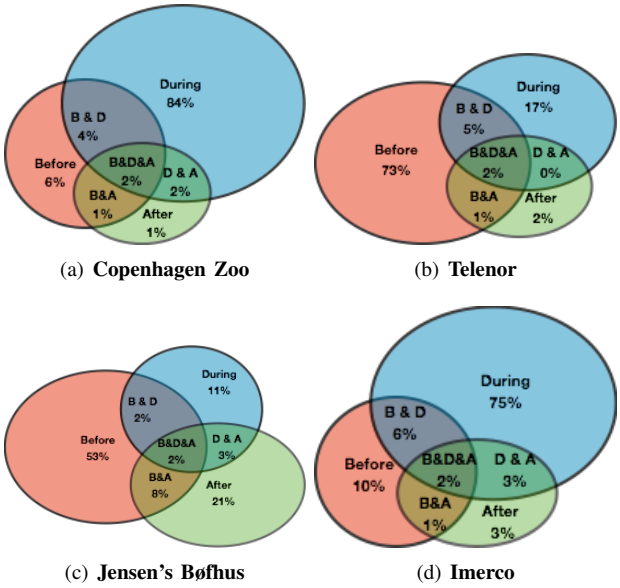
proportion of facebook users only interacted with the facebook walls of Copenhagen Zoo (86%) and Imerco (84%) during the crises period. Even for Telenor (36%) and Jensen's Bøfhus (34%), the proportion of users interacting during the crises is much higher compared to the total time period. To put it differently, SSA of actors across the time-periods of before, during and after crises confirms not only the operational definition of a social media crises but also reveals the voluminous and transient nature of user attention (that is, there many more actors interacting during the crises but they stop interacting after the crises has passed). How this change in user behaviour occurs could be a function of not only the type of social media crises it is but also the type of social media crisis communication and management strategies employed by the companies.

Figure 11 shows the temporal distribution of facebook users' likes to the artefacts (posts and comments) created by the company (facebook wall administrator). Based on associational sociology and social influence theories in social psychology, we conceptualize the action of a "facebook like" as a positive association with the artefact (facebook post or comment) and/or actor (facebook user). This type of SSA reveals the positive endorsement of the company's communication actions before, during and after the crises. As can be seen from Figure 11, surprisingly high proportion of total likes were received during the crises for the Copenhagen Zoo (84%) and Imerco (75%). This can be a structural indicator that the social media crisis might actually be a net positive for the companies concerned in terms of customer loyalty and brand parameters.

Figure 12 shows the temporal distribution of facebook users' comments to the posts created by the company. We find that the proportion of comments before and during the crises are comparable in for Jensen's Bøfhus (45% and 46%) and Telenor (29% and 66%) whereas Copenhagen Zoo (3% and 94%) and Imerco (3% and 94%) have highly skewed distribution of comments during the before and during periods of the social media crises. Since facebook doesn't have a

"dislike" button, comments are the only artefact for users to express negative associations, sentiments and expressions (also positive sentiments and expressions). Given the distribution of likes for Copenhagen Zoo and Imerco's posts and comments, the SSA of comments reveals an interesting pattern of higher likes for the company's artefacts as well as higher number of comments.

Taken together, SSA results suggest that the crisis type as well as crisis communication and management strategies employed might be different across the four cases. In order to uncover the substantive nature of the interactional patterns revealed by SSA, we conducted qualitative content analysis of the big social data corpus using two methods: (a) nethnographic analysis of the facebook walls before, during and after the crises and (b) manual sentiment analysis and topic analysis of posts and comments during the crises. These analysis help shed further analytical light on the nature of the crises and the crises communication and management strategies, if any, employed by the companies.



(a) **Copenhagen Zoo** (b) **Telenor**

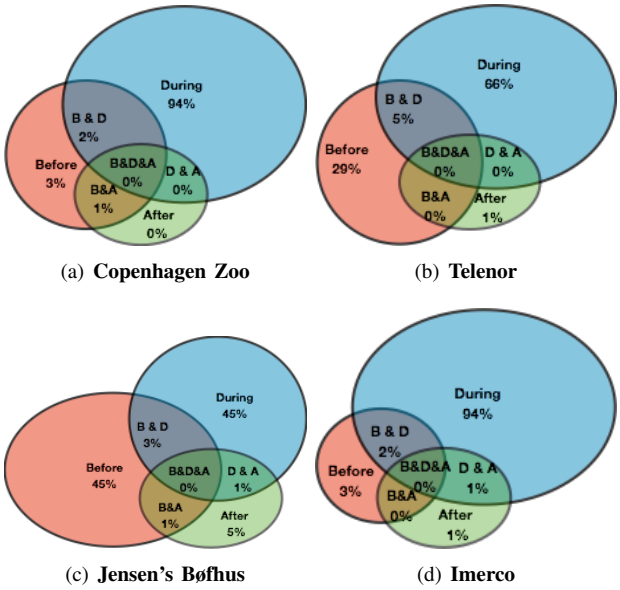
(c) **Jensen's Bøfhus** (d) **Imerco**

Figure 11. Set Analysis of actors who liked admin posts during crisis

VIII. Case Study 3: Social Set Visualizer: A Set Theoretical Approach to Big Social Data Analytics of Real-World Events

Event studies is a finance methodology to assess an impact on corporate wealth (e.g. stock prices) caused by events such as restructuring of companies, leadership change, mergers & acquisitions [73]–[75]. It has been a powerful tool since the late 1960s to assess financial impact of changes in corporate policies and used exclusively in the area of investments and accounting to examine stock price performance and the dissemination of new information [76].

While there is no unique structure for event study methodology, at a higher level of abstraction, it contains identifying three important time periods or windows. First, defining an event of interest and identifying the period over which it is active (event window), the second involves identifying the estimation period for the event (pre-event or estimation window) and the final one being identifying the post-event



(a) **Copenhagen Zoo** (b) **Telenor**

(c) **Jensen's Bøfhus** (d) **Imerco**

Figure 12. Analysis of actors who commented on admin posts during crisis

window [75]. In social set analysis of a real-world event, we have applied event study methodology to identify the three important time periods of user interactions on social media platforms: *before* (pre-event window), *during* (event window) and *after* (post-event window).

A. Methodology

Building on the formal definitions of social data from Sec. V, we further define the notion of a Facebook wall as follows,

Definition 8.1: With Social Data D, let $\mathbb{W}$ be a set of Facebook walls such that each wall $\mathsf{w} \in \mathbb{W}.\mathsf{w} \in \mathsf{R} \wedge \mathsf{r}_{\mathsf{type}}(w) = wall$.

Definition 8.2: With Social Data D, we define $\mathsf{match} \subseteq \mathsf{U} \times \mathbb{W}$ as a relation associating actors to walls as follows,

$$\mathsf{match}(u,\mathsf{w}) = \begin{cases} \top \quad \text{if } (u,\mathsf{w}) \in \rightarrow_{post} & (1) \\ \top \quad \text{if } (-,\mathsf{w}) \in \rightarrow_{post} \wedge \\ \quad (u,\mathsf{w}) \in (\rightarrow_{like} \vee \rightarrow_{share}) & (2) \\ \top \quad \text{if } \exists r.(u,r) \in \rightarrow_{post} \wedge r \rhd \mathsf{w} & (3) \\ \top \quad \text{if } \exists r.(-,r) \in \rightarrow_{post} \wedge r \rhd \mathsf{w} \wedge \\ \quad (u,r) \in (\rightarrow_{like} \vee \rightarrow_{share}) & (4) \\ \top \quad \text{if } \exists r, r'.(-,r),(u,r') \in \rightarrow_{post} \wedge \\ \quad (r \rhd \mathsf{w}) \wedge (r \rhd r') & (5) \\ \top \quad \text{if } \exists r, r'.(-,r),(-,r') \in \rightarrow_{post} \wedge \\ \quad (r \rhd \mathsf{w}) \wedge (r \rhd r') \wedge \\ \quad (u,r') \in \rightarrow_{like} & (6) \\ \bot \quad \text{otherwise} & (7) \end{cases}$$

In the def. 8.2, we define a boolean function match that keeps track whether an actor (u) interacted with a Facebook wall (w) or not. It returns true ($\top$), if the actor is the creator of the wall

(1), or if he likes the wall (2), or if he posts messages on the wall (3). Similarly making comments on posts (5) or liking or sharing (4) of posts pertaining to wall or even liking a comment will also makes the actor to belongs to a wall as formally explained in Def. 8.2.

In the analysis, the terms *user* and *actor* are used interchangeably throughout the paper without any difference in semantics.

1) Mobility of Actors across Time: As part of social set analysis, we have considered three different time frames for an event: before, during and after, which corresponds to pre-event, event and post-event timelines of the event methodology. For an event, sets containing unique actors who performed interactions *during* (U_d), *before* (U_b) and *after* (U_a) are computed. Let $\mathbf{ts}_d$, $\mathbf{ts}_b$ and $\mathbf{ts}_a$ be the sets of time spans for *during*, *before* and *after* periods respectively.

The *during* (U_d) actors set contains the actors who have either posted or commented or liked an artifact in the pre-event time period ($\mathbf{ts}_d$), can be computed as below. Let $ac \in \{post, comment\}$, then

$$\begin{aligned}\mathsf{U}_d = &\{u \mid \exists r \in \mathsf{R}.(u,r) \in \rightarrow_{post} \ \wedge \mathbf{T}(u,r,ac) \in \mathbf{ts}_d\} \ \cup \\ &\{u \mid \exists r \in \mathsf{R}.(u,r) \in \rightarrow_{like} \ \wedge \mathbf{T}(u,r,like) \in \mathbf{ts}_d\} \ \cup \\ &\{u \mid \exists r \in \mathsf{R}.(u,r) \in \rightarrow_{share} \ \wedge \mathbf{T}(u,r,share) \in \mathbf{ts}_d\}\end{aligned}$$

where $\mathbf{T}(u,r,ac)$ and $\mathbf{T}(u,r,like)$ are timestamps of the respective actions. As indicated above the set U_d contains all the unique actors that have performed at least either a *post*, or a *comment* or a *like* or a *share* on an artifact during the event period. Similarly, the unique actor sets U_b and U_a can computed easily by replacing the $\mathbf{ts}_d$ with $\mathbf{ts}_b$ and $\mathbf{ts}_a$ in the above equation, where the time stamp of the actions belongs to time spans: before ($\mathbf{ts}_b$) and after ($\mathbf{ts}_a$) the event period respectively. Finally intersections between actor sets ($\mathsf{U}_d, \mathsf{U}_b, \mathsf{U}_a$) are computed using standard set operations. As an example, the set of unique actors who have performed actions only during the event period (neither before nor after) can be computed using the principle of Venn diagram as: $\mathsf{U}_d \setminus \big((\mathsf{U}_d \cap \mathsf{U}_b) \cup (\mathsf{U}_d \cap \mathsf{U}_a)\big)$.

2) Mobility of Actors across Space: In social set analysis, mobility across space corresponds to a notion of actors interacting with different Facebook walls. Given a set of Facebook walls (W), actors mobility across space can be computed as follows.

$$\mathsf{U}^{\mathsf{W}} = \{u \mid \forall w \in \mathsf{W}.\mathsf{match}(u,w) = \top\}$$

where U^{W} is the set of actors who have interacted with all the walls in a given set of Facebook walls ($w \in \mathsf{W}$). Mobility across space is useful for analytical purposes in domains ranging from brand loyalty (actors who have visited only one wall) to social activism (actors who might be visiting many walls to express their protest over the companies).

3) Mobility of Actors across Time and Space: By combining mobility across time and space, we can compute the set of actors that have interacted within a specific time period (e.g. during event), who also have interacted with given set of walls (W) by taking intersection of two sets: $\mathsf{U}_d^{\mathsf{W}} = \mathsf{U}_d \cap \mathsf{U}^{\mathsf{W}}$.

B. Tool and Case Study

The garment industry in Bangladesh is the second-largest exporter of clothing after China, and employs more than 3 million - mainly female - workers. This is emphasized by [77] in reference to a large factory fire in Bangladesh at the 25th of November 2012 which killed 112 workers.

The garment industry in Bangladesh has rapidly grown during the past 20 years while approving of lax safety regulations and frequent accidents [78]. "Bangladesh's garment sector [..] employs forty percent of industrial workers and earns eighty percent of export revenue. Yet the majority of workers are women. They earn among the lowest wages in the world and work in appalling conditions. Trade unions and associations face brutal conditions as labour regulations are openly flouted" [79].

At April 24th, 2013, factory disasters in the Bangladeshi garment sector culminated in the largest textile industry tragedy to date with the collapse of *Rana Plaza*, a factory building in an industrial suburb of Bangladesh's capital Dhaka [80], in which more than 1100 garment workers died from the factory's collapse and subsequent fires [81]. This event has been reported by media outlets all over the world and deeply shocked many end consumers of clothing products originating from Bangladesh.

In various research publications, safety and struggles of workers in the Bangladesh garment industry have been widely discussed [79], also with special regard to ongoing protests [82], globalization-related problems [83] and ethical aspects of the factory disasters [84].

Nevertheless, the lack of publicly shown empathy by many major textile industry companies created a public outcry against perceived unethical behavior in textile industry supply chains. In many cases, this public outcry was expressed by consumers and directly addressed to the respective clothing brands, which were in the consumers' immediate reach through means of social media channels such as Facebook.

The factory disasters in Bangladesh prompted major consumer-facing textile industry brands like *H&M* and *Walmart* to join campaigns supporting textile workers' rights in Bangladesh. A more sustainable, but lagging impact is felt by the introduction of better methods of supply chain management such as social contracts in supply chains [85].

1) Methodology: Our research methodology consisted of seven steps. First, we assembled a list of real-world events with respect to the Bangladesh factory accidents. Second, we created a list of the traditional news media (print newspapers, TV and radio) reports of the real-world factory accidents in Bangladesh. Third, we reviewed the media reports and extracted a list of 11 multi-national companies (as shown in Table. VIII) that have been frequently mentioned in the traditional media reports in relation to the Bangladesh garment factory accidents. Fourth, since strategic Corporate Social Responsibility communication is conducted by companies on their Facebook pages, we extracted the full archive of the social data from the Facebook walls of the 11 companies using SODATO [8]. Fifth, we designed, developed and evaluated the Social Set Visualizer dashboard of this Facebook corpus of approximately 180 million data points. Sixth, we addressed

and answered a set of research questions using the dashboard. Seventh and last, we deployed the dashboard internally to support ongoing research by CSR researchers and practitioners.

Table VIII
OVERVIEW OF FACEBOOK DATASET OF RETAIL CLOTHING COMPANIES

Facebook Wall	Posts	Comments	Likes	Total
1) Benetton	2,411	51,156	3,760,914	3,814,481
2) Calvin Klein	12,390	44,224	3,196,564	3,253,178
3) Carrefour	3,711	18,651	79,855	102,217
4) E.C. Ingles	21,211	121,684	3,168,950	3,311,845
5) H&M	100,461	262,588	7,779,411	8,142,460
6) JC Penny	24,744	154,620	3,064,581	3,243,945
7) Mango	3,498	204,695	18,661,291	18,869,484
8) Primark	1,343	73,229	1,333,181	1,407,753
9) PVH	66	80	1,801	1,947
10) Walmart	284,523	2,147,994	44,812,653	47,245,170
11) Zara	3,136	12,437	246,294	261,867
Total:	**457,494**	**3,091,358**	**86,105,495**	**89,654,347**

2) Data Collection & Processing: The event timeline of Bangladesh factory accidents and media reports was collected through desk research including systematic searches in web and media databases. Facebook data was collected through the Social Data Analytics Tool (SODATO) [37], [38], [86]. SODATO-provided Facebook activity datasets are generated as independent files for each company's Facebook wall, and were combined into one for using them as a whole data set that can be filtered or expanded on demand. Figure 13 shows SoSeVi's system schematic for the data acquisition, processing and visualization. The general concept follows the stages of the "Big Data Value Chain" introduced by Miller and Mork [87], with steps of preparation, organization and integration of the data prior to visualization and analysis. Data preparation tasks are performed in a pre-processing step which converts all CSV files to from their character encoding UTF-16 to the more commonly used UTF-8 and handles edge cases in which the generated SODATO output lacks proper data type encapsulation. Subsequently, a data normalization phase performs sanity checks on the input data and identifies malformed data or unneeded information. Lastly, all distinct data sets are aggregated while conserving information regarding their original source in an additional variable. The aggregated data is then imported into a database management system (DBMS), from which it can be accessed for visual analytics purposes.

3) Design: In this section, the design process of the visual analytics dashboard of SoSeVi is outlined.

4) Design Goals & Objectives: The visual analytics dashboard has the following design goals.

Multidimensionality: A visual analytics dashboard consists of a mash-up of multiple visualizations which can be utilized by the user in combination to maximize efficiency. The type and size of each visualization need to be carefully evaluated.

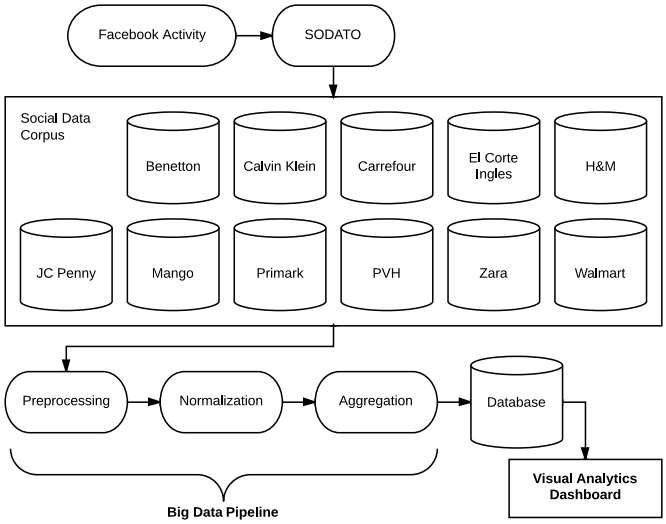


Figure 13. Big Data Acquisition Pipeline of the Social Data from Facebook used later on in the Visual Analytics Tool

Accessibility: The dashboard should be accessible as easily as possible for users. It should therefore have as few hard dependencies in terms of installed software, operating system or device type as possible.

Responsiveness: The dashboard needs to be responsive to different device types and screen sizes. It should be able to display both on a 4K display used in a conference room and a normal tablet.

Performance: A key objective for the visual analytics dashboard displays the performance in terms of both server and client side software components. As the dashboard needs to deal with large-scale data sets it should be able to process the data efficiently. In order to achieve higher performance sharing of data processing between server and client software components needs to be established. Thereby, workload may be shifted as needed and user interface waiting times are reduced.

Ease of Use: For end users, ease of use depicts an important non-functional requirement. The visual analytics dashboard should be designed in a way that enables users to work with the dashboard without any prior briefing or training on how to use it.

Extensibility: Lastly, during realization of the visual analytics dashboard, an extensible framework should be used so that future changes can be implemented with only moderate effort and without unnecessary technical hindrances.

5) Design Principles: The design of a visual analytics dashboard such as SoSeVi needs to follow a set of core principles, through which the above stated goals can be achieved. The following design principles are adopted:

Detail on Demand: The detail on demand principle strives to first present an easily graspable overview to the user, as that it can be processed visually and intellectually in short time. Only subsequently, when the user decides to, the level of detail shown in the visual analytics tool can be increased.

Ready-made Visualizations: The SoSeVi Visual Analytics dashboard is based on social media data from Facebook.

The dashboard may consist of a combination of multiple visualizations and each visualization needs to highlight unique features of the underlying social interactions between actors and artifacts. This allows the dashboard as a whole to be kept clean and organized, preventing it from becoming too complex.

User-centric Design (UCD): [88] emphasizes that in user-centric design, "the role of the designer is to facilitate the task for the user and to make sure that the user is able to make use of the product as intended and with a minimum effort to learn how to use it". When designing the interface, a focus is put on optimization of the user experience.

6) SoSeVi: Visualization Framework: The technology choice for realizing the dashboard visualizations is the D3.js Javascript-based visualization framework which uses dynamic SVG images for data visualization. D3.js constitutes a lightweight and very extendable Javascript visualization framework which can display visualizations for a multitude of browser-based clients. The flexibility provided by D3.js enables the creation of new kinds of interactive visualizations which are able to run on any device with decent processing resources including Windows, MacOS and Linux based systems with screen sizes up to 4K devices.

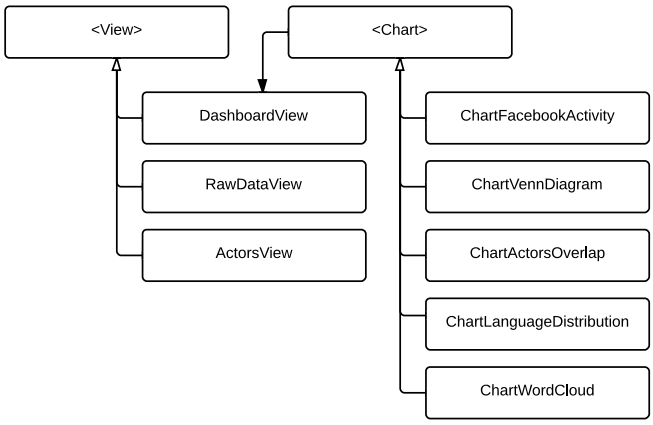


Figure 14. Software Architecture

Figure 14 presents the software architecture of SoSeVi. DashboardView is the main view of the web application which contains the SoSeVi and is initially shown to the user. RawdataView presents a detailed search interface for the Facebook activity data. Many visualizations in DashboardView refer to RawdataView in order to provide the user with further information. ActorsView presents a dedicated interface for analysis tasks related to Actor Mobility across time and space of companies' facebook walls. The visualizations of actor mobility in DashboardView refer to ActorsView in order to provide the user with further details when requested. Furthermore, ActorsView presents a handy set of tools for analysis of actor mobility and cross-postings between different time frames and Facebook walls.

7) Development of SoSeVi: Dashboard Interface: Figure 15 presents the SoSeVi dashboard and its constituent visualizations for the full dataset.

The *Facebook activity visualization* displays the social media activity on Facebook over the whole time period. It consists of a large main chart and a smaller mini chart underneath. Both charts use a line plot to display activity. The mini-chart can be used as a brush to change the time period of the data shown in the main chart. The *Actor Mobility across Space visualization* at the top right of the dashboard displays the number of different Facebook walls on which Actors have posted. For this visualization, a bar chart is used. The chart depicts the number of Actors based on the number of Facebook walls they have posted to. The *Actor Mobility across Time visualization* at the center right of the dashboard displays the number of Actors within each time period and their respective overlaps. For this visualization, an exploded Venn diagram is used which is aligned hexagonally. The *Language Distribution visualization* at the bottom right of the dashboard displays the number of social media Artifacts based on their language. For this visualization, a bar chart is used. It presents each language and the respective number of social media activities during the selected timeframe. The *Word Cloud visualization* located right beneath the Facebook Activity chart displays the results of the word frequency analysis based on conversation artifacts in the available social data. The font size of each word is determined by its overall frequency within all conversations that happened during the selected time period.

A Legend for the event timeline is placed at the very top of the dashboard between the user-driven filtering interface and the Facebook Activity visualization. It conveys information about different types of events which are part of the event timeline. In the case at hand, the event timeline is based on the Bangladesh factory disaster events, which means that the event types classified are encoded in the legend.

The user-driven filtering interface contains two components. On the left hand side, the user may input start and end dates of the timeframe to be visualized in the dashboard. Mouse or touch interactions with the input fields will reveal a hidden date picker component. This date picker enables the user to either input dates using a keyboard or specifying the day, month and year using their mouse or even a touch screen. Secondly, on the right hand side, the user may select the companies whose Facebook walls are shown in the dashboard. User interaction with the available input field can be performed in various ways. The user can directly type Facebook walls into the field, which are then displayed in the visualization. An alternative method is that the user selects an item from a drop down menu that ap-pears when the input field is focused.

To summarize, the SoSeVi big data visual analysis dashboard empowers users to use it in different ways. The dashboard adheres to the user's preferred interaction method without making any assumptions. This means tablet users may also type in their selection of the Facebook walls, or desktop users may use the Datepicker to manually select a date. The dashboard may be accessed at http://5.9.74.245:3000/, access credentials will be provided to the research community upon request.

8) Evaluation:

Benchmarking: Figure 16 displays benchmarking results of the dashboard's underlying API. The results underline

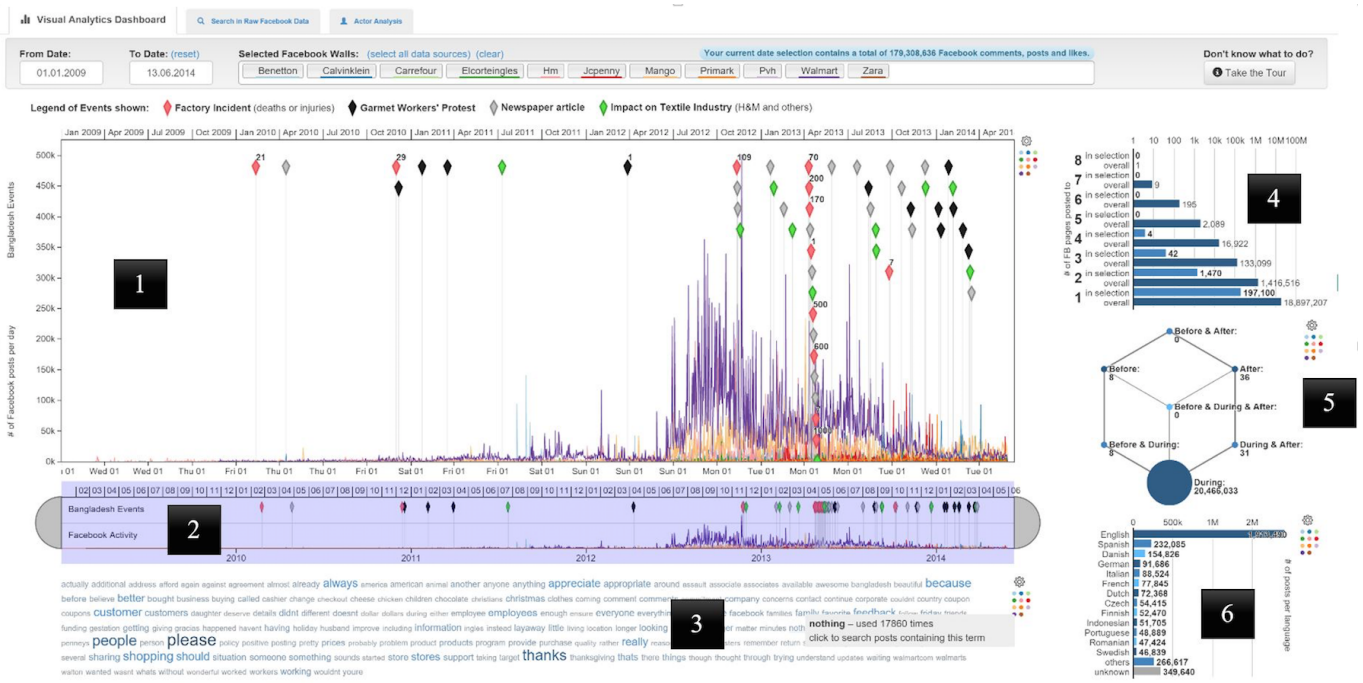


Figure 15. Social Set Visualizer: For the selected time period (see date range fields in top-left) and selected facebook walls (see colour coded selection bubble chart next to visualizations) [1] Facebook Activity Chart; [2] Timeline of Bangladesh Factory Accidents & Facebook Actions; [3] Word Cloud of Text from Posts and Comments; [4] Actor Mobility across Space (Facebook walls); [5] Actor Mobility across Time (before, during, after time-period of selection and combinations of them); and [6] Language Distribution

the varying complexity in calculating data needed for the visualizations of different event windows. According to the presented benchmark, visualizations of conversation content (*ChartLanguageDistribution* and *ChartWordCloud*) are much faster calculated and presented to the dashboard user than visualizations of actor mobility (ChartVennDiagram, ChartActorsOverlap). This can be explained by the fact that visualizations of Actor Mobility need to take each single actor into account, whereas visualizations of conversation content have access to much better speed improvements through precalculated datasets which derive from the main dataset. Due to the bad benchmark results of *ChartVennDiagram*, and a general discrepancy in performance, further optimizations are performed to the database as described further.

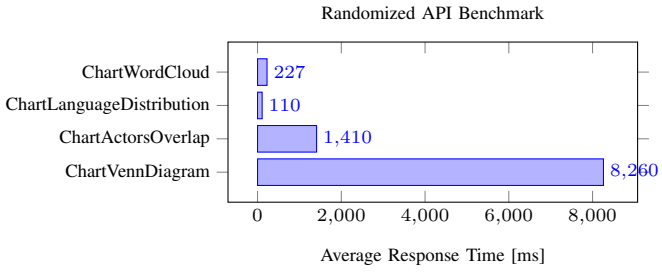


Figure 16. Performance Benchmark of four API Endpoints

Query Optimization: When using a RDBMS such as PostgreSQL in big data analytics, many opportunities for increased performance can be realized through query optimization. The systematic optimization of slow database queries is demonstrated on the visualization of Language Distribution. All optimizations are benchmarked against the initial query in order to assess their effectiveness. The benchmarking process follows a strict methodology, in which each query will be executed $n = 10$ times and query execution time is logged. Then, the average execution time is used to decide on the feasibility of the optimization at hand. If the average execution time is reduced, the optimization step will be applied to the query. The optimization process may be repeated until sufficient reduction of the average query execution time is reached. Out of all queries, the language distribution query was identified as a very slow query and therefore we have performed optimizations on it. The initial query is displayed in Listing 1. It returns 24 rows after an execution time of approximately 10 seconds, which is way slower than the users' anticipated loading time of a visual analytics dashboard. Based on the precalculation of as much data as possible and separating this data into its own database table, we optimized the performance of the query as shown in Listing 2.

Listing 1. Initial Query for Language Distribution

```
SELECT lang, COUNT(*) as count
FROM fbdata WHERE eventname != 'LIKE' AND
  "date" BETWEEN '2009-01-01' AND '2014-06-12'
  AND source in ('carrefour', 'walmart')
GROUP BY lang ORDER BY count DESC;
```

Listing 2. First Optimization of Language Distribution Query

```
CREATE TABLE fbdata_language_distribution AS
  SELECT date, source, lang, count(*) as count
  FROM fbdata GROUP BY date, source, lang
  ORDER BY date, source, lang ASC;
SELECT lang, sum(count) as count
FROM fbdata_language_distribution
WHERE "date" BETWEEN '2009-01-01' AND '2014-06-12'
  AND source in ('carrefour', 'walmart')
```

```
9  GROUP BY lang ORDER BY count DESC;
```

This performance improvement of the database query shown in listing 2 is based on the fact that the new query does not need to access the much larger $fbdata$ table, but only uses a small subset which is available in the derived table. In the second round, further optimization are performed on the query in listing 2 by creating indexes on suitable columns such as datetime and others. After creation of the indexes on the derived table, the performance of the query is increased marginally as shown in Fig. 17. A performance improvement of 300 times was realized in the first optimization, whereas the second optimization step yielded only a 1.77 times improvement.

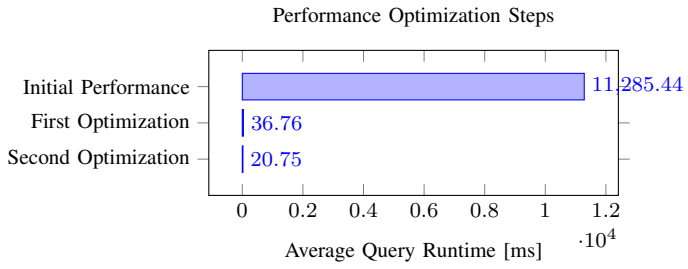


Figure 17. Three-Step Performance Optimization of the Language Distribution Visualization

9) Selected Empirical Findings from SoSeVi: Due to space restrictions, we present only a subset of the empirical findings resulting from the use of the Social Set Visualizer (SoSeVi) tool by researchers and practitioners in the field of Corporate Social Responsibility (CSR). These empirical findings demonstrate the analytical utility of our proposed set theoretical approach to big social data and our social set analysis approach to visual analytics dashboards. The following points outline some of the key issues that were investigated using the SoSeVi:

1) The global supply chain concerns with regard to Bangladesh garment factories have been expressed by Facebook users from as far back as 2009
2) With regard to Conversation analysis of big social data, the distribution of the keyword “bangladesh” across time and space of 11 different Facebook walls is proportional to the severity of fatalities in Bangladesh garment factories and peaks for the Rana Plaza disaster that killed more than 1100 factory workers.
3) With regard to Interaction analysis of big social data, in terms of actors, SoSeVi helped identify the most influential negative critics as well as positive advocates for each of the 11 companies before, during, and after the maximum accident density time period
4) There are many instances of authentic displays of support and expressions of empathy from Facebook users as well as robotic incidents of slacktivism
5) Surpisingly, majority usage the keyword *please* was with respect to opening of new stores in the case of H&M
6) Protestors and activists employed different social media strategies on the different Facebook walls of companies but with little evidence for social influence (in terms of the number of likes and comments on their posts)
7) Companies followed not only different CSR strategies but also different social media strategies before, during and after the Bangladesh garment factory accidents. Further, companies adopted different crises communication and management strategies yielding different outcomes
8) For almost all of the accidents, a majority of the users are posting during the news cycle, e.g. the coverage of the event through traditional media outlets, and they do not return to the Facebook walls again. This emphasizes that social media engagement during factory accidents is episodic and burst-y with little overlap to the *business-as-usual* period before or after the accident.

10) Reflections on the IT-Artifact: Computational social science research has reached a point where social media activity is ubiquitous yet hard to collect and analyze in domain-specific ways (with the notable exception of epidemiology). In conjunction with complex event timelines as depicted by the Bangladesh garment factory disasters, the data at hand presents numerous opportunities for attaining deep insights. In this context, visual analytics present the means of reaching those insights to many users with different backgrounds, both experts and novices alike. The novel implementation of the present Social Set Visualizer (SoSeVi) dashboard showcases that the creation of visual analytics software, which meets the high technical, analytical and user experience requirements of present-day computing, is viable (and can be achieved by an academic research group with limited resources). Furthermore, the developed IT artifact leverages open-source visual analytics frameworks to maximum extent in order to achieve a pure implementation of important concepts in visual analytics.

IX. Discussion

In this paper we have presented an new approach for analysis of big social data using a conceptual model of social data, a set-theoretical formalisation of the conceptual model and an analytical framework called ”Social Set Analysis”. The set-theoretical formalization of the conceptual model provides the necessary abstraction to comprehend the complex and complicated interactional scenarios and conversational contexts of big social data. Further, the formal model informed the schematic model of the software application and helped realise the abstract ideas from the conceptual model into the analytical framework for Social Set Analysis. We now briefly discuss the methods of and findings from the three illustrative case studies.

A. Case Study 1: Fuzzy-Set Based Sentiment Analysis

In the first case study (Sec. VI), we have presented an integrated modelling approach for analysis of big social data with the sentiment analysis based on the Fuzzy set theory. We have presented a method for profiling of artifacts and actors and applied this technique to the analysis of big social data collected from Facebook page of the fast fashion company, H&M. Regarding formal modelling of temporal dimensions of social media interactions, we are currently developing a

hybrid approach by constructing crisp sets as well as fuzzy sets. For example, given an event of analytical interest such as a marketing campaign, we construct crisp sets of sentiment categories (positive, negative, & neutral) for actors and artefacts and fuzzy sets of the interactional time-periods (before_event, during_event, and after_event). This allows us to model and analyze the different user characteristics, behaviours, and dynamics within the intersections and unions of the temporal categories of before-the-event, during-the-event, and after-the-event at analyst determined fuzzy set membership criteria for sentiment categories.

We acknowledge that many works exist in fuzzy sentiment analysis and social networks and we have cited relevant papers in the related work section. But as stated before, our approach primarily differs from the current approaches of social network analysis based on relational sociology. Our approach is based on associational sociology, where we focussed on finding ”association-ship” among actors and artifacts, based on set theoretical approach, rather than only focussing on the relationship between the actors. Our approach of ”associational sociology” is drawn from Bruno Latour’s ([25]) term ”sociology of associations”. We postulate that Set Theory in general and Fuzzy Set Theory is well-suited from sociological and mathematical standpoints to model human associations [27]. Beyond the immediate social network and particularly on large scale social media platforms such as Facebook, twitter and Tencent QQ, we believe that this fundamental change in the foundational mathematical logic of the formal model of social data from graphs to sets will generate new insights. This paper is a first attempt to articulate such an alternate integrated approach across the theoretical, conceptual, formal and computational realms.

B. Case Study 2: Social Set Analysis Of Corporate Social Media Crises On Facebook

In the second case study (Sec. VII), we first proposed a set-theoretical formal method for social set analysis drawn from the event-study framework to investiage corporate social media cries on Facebook. The proposed method was then applied to big social data for four different social media crises. Event studies is a finance methodology to assess an impact on corporate wealth (e.g. stock prices) due to events such as restructuring of companies, leadership change, mergers & acquisitions [73]–[75]. It has been a powerful tool since late 1960s to assess financial impact of changes and continues to be used extensively to examine stock price performance and the dissemination of new information [76]. While there is no unique structure for event study methodology, at an higher level of abstraction, it contains identifying three important time periods or windows. First, defining an event of interest and identify the period over which it is active (event window), the second involves identifying the estimation period for the event (pre-event or estimation window) and the final one being identifying the post-event window [75]. In social set analysis of social media crises, we have adopted the event study methodology to identify the three important time periods of user interactions on social media platforms: *before* (pre-event window), *during* (event window) and *after* (post-event window). SSA results showed the voluminous but also transient nature of interactions during the social media crises and a diversity of aggregate user behavioural patterns. SSA combined with netnography and content analysis in terms of sentiment analysis and topic discovery revealed the different strategies employed by the organizations to manage the crises and their outcomes.

C. Case Study 3: Social Set Visualizer: A Set Theoretical Approach To Big Social Data Analytics Of Real-World Events

In the third case study (Sec. VIII), we designed, developed and evaluated a visual analytics tool, SoSeVi(Social Set Visualiser). SoSeVi incorporated new set theoertical visualizations of big social data in terms of actor mobility across space (wall crossings) and actor mobility across time (before, during and after). SoSeVi leverages open-source visual analytics frameworks to maximum extent in order to achieve a pure implementation of important concepts in visual analytics such as the detail on demand principle. A thorough evaluation showcased the effectiveness of the tool’s approach on visual analytics. Both the client- and a server-side components of the Visual Analytics Dashboard present performance at par with commercial tools, and can seamlessly be used under many circumstances. Additionally, the results of the user study performed indicate that the presented Visual Analytics dashboard combines a high ease of use with the ability of performing many different interactive analyses on a large dataset. Moreover, the Visual Analytics tool put forward may be utilized through any modern browser on a multitude of different devices and screen sizes, with backend response times as low as in the hundreds of milliseconds. Complementing benchmarks of the database optimizations applied to the Visual Analytics dashboard in real-world deployments showcase a good performance and satisfactory handling of large amounts of social data.

D. Reflections on the Set-Theoretical Approach

We now briefly discuss the current adoption of and future prospects for the set-theortical approach with regard to Social Science, Computer Science, and Computational Social Science.

1) Set Theory and Social Science: Recent advancements in set theory and readily available software have enabled social science researchers to bridge the variable-centered quantitative and case-based qualitative methodological paradigms in order to analyze multi-dimensional associations beyond the linearity assumptions, aggregate effects, unicausal reduction, and case specificity. In the social sciences, application of set theory has seen a dramatic increase over the last decade which can be attributed to the method called âQualitative Comparative Analysisâ [89] developed by the political scientist Charles Ragin [90], [91]. Qualitative Comparative Analysis (QCA) models causal relations as subset or superset relations corresponding to necessity and sufficiency conditions. QCA aims to derive casually complex patterns in terms of equifinality, conjunctural causation and asymmetry [90], [92], [93].

Although developed initially by Ragin [91] for qualitative case study researchers (medium sample size of $n < 90$), the proponents and supporters of QCA have argued about its unique advantages over regression-based approaches [93], [94] and its application for analysis of large-N datasets. In the adoption of set theoretical methods in social sciences [89] three variants of QCA methodology have surfaced: crisp-set QCA(CsQCA), fuzzy-set QCA (fsQCA) [90] and multi-set QCA (MvQCA) [93] with a number of software tools supporting set-theoretical social science research (e.g. R packages like QCA and QCAPro, fs/QCA, Tosmana).

2) Set Theory and Computer Science: In order to further systematic research on set-theoretical algorithms, data structures and programs in Computer Science, we envision *Computational Set Analysis* as a research program. In this regard, the SetVR workshop series[6] augurs well for the formalisation and computational implementation of set-theoretical reasoning and visualisations. In terms of visual analytics, recent advancements with regard to set intersections include the *Upset* project [95] on the visualizations of set intersections based on innovative approaches to combination matrices and the *Euler Diagrams* project on creating area-proportional Euler diagrams using ellipses instead of the traditionally used circles [96].

3) Set Theory and Computational Social Science: As discussed in the Conceptual Framework section, set-theoretical approaches big social data analytics hold several advantages in terms of modelling the implicit vagueness of many social science concepts and combining the strengths and addressing the weakness of variable vs. case based empirical approaches in social science research. For example, automated sentiment annotation of social data artifacts based on computational linguistics methods such as supervised machine learning produce both classifications of tokens into types (such as positive, negative and neutral) as well as probabilistic estimates. As we have demonstrated in Case Study 1, these classifications and probabilities can be modelled using Crisp and Fuzzy Set theories respectively and analyzed to reveal historical developmental patterns as well as overlapping categories. Practical implications from the analysis could help inform an organization to assess the size of the different actor sets (sub-communities) such as entirely positive, partially positive, entirely negative etc. Investigating the absolute and relative size of entirely negative conversations might enable the organization to identify the underlying customer service issues and/or content problems. Similarly, knowing the absolute and relative number of social media users that exclusively express positive sentiments towards the organization helps identify and nurture the advocacy group.

E. Limitations

One of this paper's limitations is that we do not present domain-specific social science findings in terms of visual analytics, crisis communication, crisis management, labor rights, industrial safety and/or corporate social responsibility. That said, first attempts at domain-specific empirical findings of the set-theoretical approach can be found in [72], [97]. A second limitation is the lack of exposition of the full range of set theory beyond the classical crisp sets and fuzzy sets discussed in the paper (for example: Rough sets, Random sets, Bayesian sets). A third and final limitation is the limited space given to the computational aspects of the visuala analytics tool, SoSeVi.

[6]SetVR workshop: https://sites.google.com/site/setvr2kn/current-workshop

F. Future Research

Current and planned future work in our computational social sciences laboratory is addressing some of the theoretical limitations identified above. In particular, we are exploring novel set-theoretical visualisations of large number of set interestions and to indicate set migrations of actors across space and time with a focus on dynamic set composition and decomposition. In terms of formal models and analytical methods, we are extending Social Set Analysus to include Rough and Random sets. Furthermore, we plan to release a software library for "Social Set Analysis" that will allow researchers and practitioners to easily integrate set-theoretical analytics into their Big Data Analytics workbenches.

X. Conclusion

In conclusion, one of the contributions of this paper is to demonstrate the suitability and effectiveness of Social Set Analysis for conceptualizing, formalizing and analyzing big social data from content-driven social media platforms like Facebook for event studies such as unexpected crises and/or coordinated marketing campaigns. Computational social science research has reached a point where social media activity is ubiquitous, yet hard to collect and analyze in domain-specific ways (with the notable exception of epidemiology). In conjunction with complex event timelines as depicted by the Bangladesh garment factory disasters, user actions on various organisation's Facebook walls, Big Social Data presents numerous opportunities for attaining deep insights. As illustrated by the three case studies above, SSA covers the range of prescriptive, visual, and descriptive analytics. Taken together, the three demonstrative case studies illustrate the viability of Social Set Analysis as a holistic approach to Computational Social Science in general and Big Data Analytics in particular.

As part of future work, we would like to extend the Fuzzy Set Theoretical formal model to encompass modelling of networks of groups and friends of users in an online social media platform. We also have plans to extend the Fuzzy Set methods and techniques to other kinds of socio-technical interactions and further develop our abstract formal model. Modelling social concepts in general involves fuzziness and we would like to use Fuzzy set theory to model fuzzy behaviour in the social data.

Acknowledgments

We thank the members of the Computational Social Science Laboratory (http://cssl.cbs.dk) for their valuable feedback. Thanks to the master theses and course project students that have helped in the design, development, use and evalaution of the methods and tools.

The authors were partially supported by the project *Big Social Data Analytics: Branding Algorithms, Predictive Models, and Dashboards* funded by *Industriens Fond* (The Danish Industry Foundation). Any opinions, findings, interpretations, conclusions or recommendations expressed in this paper are those of its authors and do not represent the views of the Industriens Fond (The Danish Industry Foundation).

REFERENCES

[1] R. E. Montalvo, "Social media management," *International Journal of Management & Information Systems (IJMIS)*, vol. 15, no. 3, pp. 91–96, 2011. 1

[2] C. Vollmer and G. Precourt, *Always on: Advertising, marketing, and media in an era of consumer control*. McGraw Hill Professional, 2008. 1, 9

[3] A. McAfee, *Enterprise 2.0: New collaborative tools for your organization's toughest challenges*. Harvard Business Press, 2009. 1

[4] R. Vatrapu, "Understanding social business." in *Emerging Dimensions of Technology Management*. Springer, 2013, pp. 147–158. 1, 9

[5] W. S. Cleveland, "Data science: an action plan for expanding the technical areas of the field of statistics," *International Statistical Review*, vol. 69, no. 1, pp. 21–26, 2001. [Online]. Available: http://dx.doi.org/10.1111/j.1751-5823.2001.tb00477.x 1

[6] M. Loukides, *What Is Data Science?* O'Reilly Media, 2012. 1

[7] N. Ohsumi, "From data analysis to data science," in *Data Analysis, Classification, and Related Methods*. Springer Berlin Heidelberg, 2000, pp. 329–334. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-59789-3_52 1

[8] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barab?si, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne, "Computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009. 1

[9] J. Sterne, *Social media metrics: How to measure and optimize your marketing investment*. John Wiley & Sons, 2010. 1, 3

[10] M. Sponder, *Social media analytics: Effective tools for building, interpreting, and using metrics*. McGraw Hill Professional, 2011. 1, 3

[11] Z. Tufekci, "Big questions for social media big data: Representativeness, validity and other methodological pitfalls," *arXiv preprint arXiv:1403.7400*, 2014. 1, 2

[12] C. Cioffi-Revilla, *Introduction to Computational Social Science: Principles and Applications*. Springer Science & Business Media, 2013. 1, 2

[13] R. Vatrapu, R. R. Mukkamala, and A. Hussain, "Towards a set theoretical approach to big social data analytics: Concepts, methods, tools, and empirical findings," in *proceedings of International Conference on Social Media & Society (#SMSociety14)*, 2014. 2

[14] S. Wasserman and K. Faust, *Social network analysis: Methods and applications (Vol. 8)*. Cambridge university press, 1994. 2

[15] M. Emirbayer, "Manifesto for a relational sociology," *The American Journal of Sociology*, vol. 103(2), pp. 281–317, 1997. 2

[16] R. Vatrapu, A. Hussain, N. B. Lassen, R. R. Mukkamala, B. Flesch, and R. Madsen, "Social set analysis: four demonstrative case studies," in *Proceedings of the 2015 International Conference on Social Media & Society*. ACM, 2015, p. 3. 2

[17] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, "Network analysis in the social sciences," *Science*, vol. 323(5916), pp. 892–895, 2009. 2

[18] J. L. Gross and J. Yellen, *Graph theory and its applications*. CRC press, 2005. 2

[19] M. S. Mizruchi, "Social network analysis: Recent achievements and current controversies," *Acta sociologica*, vol. 37, no. 4, pp. 329–343, 1994. 2

[20] R. Vatrapu, R. R. Mukkamala, and A. Hussain, "Set theoretical approach to big social data analytics: concepts, methods, tools, and findings," in *Computational Social Science workshop at European Conference on Complex Systems (ECSS-2014)*, 2014. 2

[21] F. Cusset, *French Theory: How Foucault, Derrida, Deleuze, & Co. Transformed the Intellectual Life of the United States*. U of Minnesota Press, 2008. 2

[22] I. Hacking, *The social construction of what?* Harvard university press, 1999. 2

[23] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Information, communication & society*, vol. 15, no. 5, pp. 662–679, 2012. 2

[24] J. Lovett, *Social media metrics secrets*. John Wiley & Sons, 2011, vol. 159. 3

[25] B. Latour, *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press, USA, 2005. 3, 24

[26] C. C. Ragin, *Fuzzy-set social science*. University of Chicago Press, 2000. 3, 5

[27] R. R. Mukkamala, A. Hussain, and R. Vatrapu, "Towards a set theoretical approach to big data analytics," in *proceedings of 3rd International Congress on Big Data (IEEE BigData 2014)*, June 2014, http://www.itu.dk/people/rao/pubs_accepted/2014_IEEE-BigData-socialdata-set-theory.pdf. 3, 6, 24

[28] M. J. Smithson and J. Verkuilen, *Fuzzy Set Theory : Applications in the Social Sciences (Quantitative Applications in the Social Sciences)*. SAGE Publications, Feb. 2006. [Online]. Available: http://www.worldcat.org/isbn/076192986X 3, 5

[29] A. Kechris, *Classical descriptive set theory*. Springer Science & Business Media, 2012, vol. 156. 3, 5

[30] R. K. Vatrapu, "Technological intersubjectivity and appropriation of affordances in computer supported collaboration," Ph.D. dissertation, University of Hawaii at Manoa, USA, 2007, aAI3302125. 3, 4

[31] ——, "Towards a theory of socio-technical interactions," in *Learning in the Synergy of Multiple Disciplines*. Springer, 2009, pp. 694–699. 3, 4

[32] ——, "Explaining culture: An outline of a theory of socio-technical interactions," in *Proceedings of the 3rd International Conference on Intercultural Collaboration*, ser. ICIC '10. New York, NY, USA: ACM, 2010, pp. 111–120. 3, 4, 14

[33] J. J. Gibson, *The ecological approach to visual perception*. Boston: Houghton Mifflin, 1979. 3

[34] A. Noë, *Action in perception*. MIT press, 2004. 3

[35] A. Schuetz, *The phenomenology of the social world*. Northwestern University Press, 1967. 4

[36] H. Garfinkel, *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall, 1967. 4

[37] R. Vatrapu, A. Hussain, D. Hardt, and Z. Jaffari, "Social data analytics tool: A demonstrative case study of methodology and software." in *Analysing Social Media Data and Web Networks*, R. Gibson, Ed. Palgrave Macmillan, 2014. 4, 5, 14, 20

[38] A. Hussain and R. Vatrapu, "Social data analytics tool (sodato)," in *DESRIST-2014 Conference (in press)*, ser. Lecture Notes in Computer Science (LNCS). Springer, 2014. 5, 12, 20

[39] M. Kryszkiewicz, "Rough set approach to incomplete information systems," *Information sciences*, vol. 112, no. 1, pp. 39–49, 1998. 5

[40] N. M. Tichy, M. L. Tushman, and C. Fombrun, "Social network analysis for organizations," *The Academy of Management Review*, vol. 4, no. 4, October 1979. 5

[41] D. Krackhardt, "Cognitive social structures," *Social Networks*, vol. 9, no. 2, pp. 109–134, Jun. 1987. 5

[42] J. Zhan and X. Fang, "Social computing: the state of the art," *International Journal of Social Computing and Cyber-Physical Systems*, vol. 1, no. 1, pp. 1–12, 01 2011. 5

[43] J. Karikoski and M. Nelimarkka, "Measuring social relations with multiple datasets," *IJSCCPS*, vol. 1, no. 1, pp. 98–113, 2011. 5

[44] J. Sabater and C. Sierra, "Reputation and social network analysis in multi-agent systems," in *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1*, ser. AAMAS '02. New York, NY, USA: ACM, 2002, pp. 475–482. 5

[45] M. Goldberg, S. Kelley, M. Magdon-Ismail, K. Mertsalov, and A. Wallace, "Finding overlapping communities in social networks," in *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, 2010, pp. 104–113. 5

[46] O. Macindoe and W. Richards, "Comparing networks using their fine structure," *International Journal of Social Computing and Cyber-Physical Systems*, vol. 1, no. 1, 2011. 5

[47] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008. 5

[48] C. R. Fink, D. S. Chou, J. J. Kopecky, and A. J. Llorens, "Coarse-and fine-grained sentiment analysis of social media text," *Johns Hopkins APL Technical Digest*, vol. 30, no. 1, pp. 22–30, 2011. 5

[49] M. Grassi, E. Cambria, A. Hussain, and F. Piazza, "Sentic web: A new paradigm for managing social media affective information," *Cognitive Computation*, vol. 3, no. 3, pp. 480–489, 2011. 5

[50] T. Nguyen, D. Phung, B. Adams, and S. Venkatesh, "Prediction of age, sentiment, and connectivity from social media text," in *Web Information System Engineering–WISE 2011*. Springer, 2011, pp. 227–240. 5

[51] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011. 5

[52] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis," *Computational linguistics*, vol. 35, no. 3, pp. 399–433, 2009. 5

[53] S. Asur and B. Huberman, "Predicting the future with social media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 1, 2010, pp. 492–499. 5, 6

[54] H. Chen, P. De, Y. Hu, and B.-H. Hwang, "Sentiment revealed in social media and its effect on the stock market," in *Statistical Signal Processing Workshop (SSP), 2011 IEEE*. IEEE, 2011, pp. 25–28. 5

[55] D. Hardt and J. Wulff, "What is the meaning of 5 *'s? An investigation of the expression and rating of sentiment," in *Proceedings of KONVENS 2012*, J. Jancsary, Ed. OGAI, September 2012, pp. 319–326, pATHOS 2012 workshop. 5

[56] S. P. Robertson, "Changes in referents and emotions over time in election-related social networking dialog," in *System Sciences (HICSS), 2011 44th Hawaii International Conference on*. IEEE, 2011, pp. 1–9. 6

[57] M. Salathé and S. Khandelwal, "Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control," *PLoS computational biology*, vol. 7, no. 10, p. e1002199, 2011. 6

[58] T. Menezes, C. Roth, and J.-P. Cointet, "Finding the semantic-level precursors on a blog network." *IJSCCPS*, vol. 1, no. 2, pp. 115–134, 2011. 6

[59] R. R. Mukkamala, A. Hussain, and R. Vatrapu, "Towards a formal model of social data," IT University of Copenhagen, Denmark, IT University Technical Report Series TR-2013-169, November 2013. 6

[60] McDonalds, "Mcdonalds facebook post," https://www.facebook.com/McDonalds/photos/a.10150151878945584.414818.10150097174480584/10156340092750584/?type=3. 8

[61] H. Li and J. D. Leckenby, *Internet Advertising: Theory and Research*. Psychology Press; 2 edition, 2007, ch. Internet Advertising Formats and Effectiveness. 9

[62] C. P. Haugtvedt, P. M. Herr, and F. R. Kardes, *Handbook of consumer psychology*. Psychology Press, 2012. 9

[63] K. Kunst and R. Vatrapu, "Towards a theory of socially shared consumption: Literature review, taxonomy and research agenda," in *Proceedings of the European Conference on Information Systems (ECIS), Tel Aviv, Israel*, 2014. 9

[64] T. E. Barry, "The development of the hierarchy of effects: An historical perspective," *Current issues and Research in Advertising*, vol. 10, no. 1-2, pp. 251–295, 1987. 9

[65] R. J. Lavidge and G. A. Steiner, "A model for predictive measurements of advertising effectiveness." *Journal of marketing*, vol. 25, 1961. 9

[66] T. Veblen, *The theory of the leisure class; an economic study of institutions*. Oxford University Press, 1899(2009). 9

[67] H.-J. Zimmermann, "Fuzzy set theory," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 3, pp. 317–332, 2010. [Online]. Available: http://dx.doi.org/10.1002/wics.82 9

[68] Google-Inc, "Google prediction api," September 2012, https://developers.google.com/prediction/. 12, 13

[69] W. T. Coombs, "Choosing the right words the development of guidelines for the selection of the ÒappropriateÓ crisis-response strategies," *Management Communication Quarterly*, vol. 8, no. 4, pp. 447–476, 1995. 14

[70] ——, *Ongoing crisis communication: Planning, managing, and responding*. Sage Publications, 2014. 14

[71] C. Zimmerman, Y. Chen, D. Hardt, and R. Vatrapu, "Marius, the giraffe: a comparative informatics case study of linguistic features of the social media discourse," in *Procs. of conference on Collaboration across boundaries: culture, distance & technology*. ACM, 2014, pp. 131–140. 14

[72] R. R. Mukkamala, J. I. Sorensen, A. Hussain, and R. Vatrapu, "Detecting corporate social media crises on facebook using social set analysis," in *Big Data (BigData Congress), 2015 IEEE International Congress on*. IEEE, 2015, pp. 745–748. 16, 17, 25

[73] P. Bromiley, M. Govekar, and A. Marcus, "On using event-study methodology in strategic management research," *Technovation*, vol. 8, no. 1, pp. 25–42, 1988. 18, 24

[74] A. McWilliams and D. Siegel, "Event studies in management research: Theoretical and empirical issues," *Academy of management journal*, vol. 40, no. 3, pp. 626–657, 1997. 18, 24

[75] A. C. MacKinlay, "Event studies in economics and finance," *Journal of economic literature*, pp. 13–39, 1997. 18, 24

[76] J. Binder, "The event study methodology since 1969," *Review of quantitative Finance and Accounting*, vol. 11, no. 2, pp. 111–137, 1998. 18, 24

[77] V. Bajaj, "Fatal fire in bangladesh highlights the dangers facing garment workers," *New York Times*, vol. 25, 2012. 19

[78] H. Sato, "Cournot competition and reduction of corruption to prevent garment factory fires in bangladesh," *Advances in Management and Applied Economics*, vol. 4, no. 4, pp. 17–20, August 2014. 19

[79] P. Khanna, "Making labour voices heard during an industrial crisis: Workers' struggles in the bangladesh garment industry," *Labour, Capital and Society/Travail, capital et société*, pp. 106–129, 2011. 19

[80] J. A. Manik, J. Yardley, and B. DHAKA, "Building collapse in bangladesh leaves scores dead," *NY TIMES (Apr. 24, 2013), http://www.nytimes. com/2013/04/25/world/asia/bangladesh-buildingcollapse. html*, 2013. 19

[81] J. Burke, "Bangladeshi factory collapse leaves trail of shattered lives," *The Guardian*, 2013. 19

[82] S. A. Himi and A. Rahman, "Workers unrest in garment industries in bangladesh: An exploratory study," *Journal of Organization and Human Behaviour*, vol. 2, no. 3, pp. 49–55, 2013. 19

[83] S. Rahman, *Broken Promises of Globalization: The Case of the Bangladesh Garment Industry*. Lexington Books, 2013. 19

[84] K. L. Stewart, "An ethical analysis of the high cost of low-priced clothing," *NABET*, p. 128, 2013. 19

[85] M. A. Islam, C. Deegan *et al.*, "Social audits and multinational company supply chain: A study of rituals of social audits in the bangladesh garment industry," *Available at SSRN 2466129*, 2014. 19

[86] A. Hussain and R. Vatrapu, "Social data analytics tool: Design, development, and demonstrative case studies," in *Enterprise Distributed Object Computing Conference Workshops and Demonstrations (EDOCW), 2014 IEEE 18th International*, Sept 2014, pp. 414–417. 20

[87] H. G. Miller and P. Mork, "From data to decisions: a value chain for big data," *IT Professional*, vol. 15, no. 1, pp. 57–59, 2013. 20

[88] C. Abras, D. Maloney-Krichmar, and J. Preece, "User-centered design," *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications*, vol. 37, no. 4, pp. 445–56, 2004. 21

[89] A. Thiem and A. Dusa, *Qualitative comparative analysis with R: A user?s guide*. Springer Science & Business Media, 2012, vol. 5. 24, 25

[90] C. C. Ragin, *Redesigning social inquiry: Fuzzy sets and beyond*. Wiley Online Library, 2008, vol. 240. 24, 25

[91] ——, *The comparative method: Moving beyond qualitative and quantitative strategies*. University of California Press, 1987. 24, 25

[92] P. C. Fiss, "A set-theoretic approach to organizational configurations," *Academy of management review*, vol. 32, no. 4, pp. 1180–1198, 2007. 24

[93] C. Wagemann and C. Q. Schneider, "Qualitative comparative analysis (QCA) and fuzzy-sets: Agenda for a research approach and a data analysis technique," *Comparative Sociology*, vol. 9, no. 3, pp. 376–396, 2010. 24, 25

[94] P. Emmenegger, D. Schraff, and A. Walter, "QCA, the truth table analysis and large-n survey data: The benefits of calibration and the importance of robustness tests," in *Paper presented at the 2nd International QCA Expert Workshop, Zurich, Switzerland*. Citeseer, 2014, Conference Proceedings. 25

[95] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister, "Upset: Visualization of intersecting sets," *IEEE Transactions on Visualization and Computer Graphics (IEEE InfoVis '14)*, 2014, live Demo: http://vcg.github.io/upset. 25

[96] L. Micallef and P. Rodgers, "euler ape: Drawing area-proportional 3-venn diagrams using ellipses," *PloS one*, vol. 9, no. 7, p. e101717, 2014. 25

[97] R. R. Mukkamala, J. I. Sorensen, A. Hussain, and R. Vatrapu, "Social set analysis of corporate social media crises on facebook," in *Enterprise Distributed Object Computing Conference (EDOC), 2015 IEEE 19th International*. IEEE, 2015, pp. 112–121. 25

Ravi Vatrapu is a professor of human computer interaction at the Department of IT Management, Copenhagen Business School; professor of applied computing at the Westerdals Oslo School of Arts Communication and Technology; and director of the Computational Social Science Laboratory (http://cssl.cbs.dk). Prof. Vatrapu's current research focus is on big social data analytics. Based on the enactive approach to the philosophy of mind and phenomenological approach to sociology and the mathematics of classical, fuzzy and rough set theories, his current research program seeks to design, develop and evaluate a new holistic approach to computational social science, Social Set Analytics (SSA). SSA consists of novel formal models, predictive methods and visual analytics tools for big social data. Prof. Vatrapu holds a Doctor of Philosophy (PhD) degree in Communication and Information Sciences from the University of Hawaii at Manoa, a Master of Science (M.Sc) in Computer Science and Applications from Virginia Tech, and a Bachelor of Technology in Computer Science and Systems Engineering from Andhra University.

Raghava Rao Mukkamala is an assistant professor of Computational Social Science at the Department of IT Management, Copenhagen Business School; external lecturer of applied computing at the Westerdals Oslo School of Arts Communication and Technology; and co-director of the Computational Social Science Laboratory (http://cssl.cbs.dk). Raghava's current research focus is on interdisciplinary approach to big data analytics. Combining formal/mathematical modelling approaches with data/text mining techniques and machine learning methodologies, his current research program seeks to develop new algorithms and techniques for big data analytics such as Social Set Analytics. Raghava holds a PhD degree in Computer Science and a M.Sc degree in Information Technology, both from IT University of Copenhagen, Denmark and a Bachelor of Technology degree from Jawaharlal Nehru Technological University, India. Before moving to research, Raghava has many of years of programming and IT development experience from Danish IT industry.

Abid Hussain is an assistant professor of Computational Social Science at the Department of IT Management, Copenhagen Business School and Associate Researcher at the Computational Social Science Laboratory (http://cssl.cbs.dk). Abid's research focus is on the design, development and evaluation of design principles and design patterns for the systematic collection, storage, retrieval and processing of big social data. He is the lead researcher and developer of the Social Data Analytics Tool (www.sodato.net), the first research-based big social data analytics tool for Facebook. He has more than ten years of software development experience in the IT industry where he has served as the system architect and lead developer on software teams ranging upto 40 members. He holds a Master of Science in Software Development from the IT University of Copenhagen, Denmark; a Graduate Diploma in Information Systems Management from the Central Queensland University, Australia; and a Diploma in Information Technology and Software Development from Holmesglen Institute, Australia.

Benjamin Flesch is a PhD Fellow of Computational Social Science at the Department of IT Management, Copenhagen Business School and research fellow at the Computational Social Science Laboratory (http://cssl.cbs.dk). His research aims to formulate and evaluate a new field of study, *Computational Set Analysis* in general and Computational Set Visualisations in particular. His PhD project aims to design, develop and evaluate Social Set Visualiser (SoSeVi) based on set-theoretical approach to computational social science. He holds a Master of Science in Business Administration and Information Systems from Copenhagen Business School, Denmark and Master of Science in Business Administration from University of Mannheim, Germany.

Publication II

Social Set Visualizer: Demonstration of Methodology and Software

---

Benjamin Flesch, Abid Hussain and Ravi Vatrapu. *Social Set Visualizer: Demonstration of Methodology and Software*. In 2015 IEEE 19th International Enterprise Distributed Object Computing Workshop, pages 148–151, Sept 2015

Social Set Visualizer: Demonstration of Methodology and Software

Benjamin Flesch[1]
[1]Computational Social Science Laboratory
Department of IT Management
Copenhagen Business School
bfitservice@gmail.com

Abid Hussain[1]
[1]Computational Social Science Laboratory
Department of IT Management
Copenhagen Business School
ah.itm@cbs.dk

Ravi Vatrapu [1,2]
[2]Mobile Technology Laboratory
Faculty of Technology
Westerdals Olso School of ACT
vatrapu@cbs.dk

Abstract

This paper presents a state-of-the art visual analytics dashboard, Social Set Visualizer (SoSeVi), of approximately 90 million Facebook actions from 11 different companies that have been mentioned in the traditional media in relation to garment factory accidents in Bangladesh. The enterprise application domain for the dashboard is Corporate Social Responsibility (CSR) and the targeted end-users are CSR researchers and practitioners. The design of the dashboard was based on the "social set analytics" approach to computational social science [1]. The development of the dashboard involved cutting-edge open source visual analytics libraries from D3.js and creation of new visualizations (actor mobility across time, conversational comets etc). Evaluation of the dashboard consisting of technical testing, usability testing, and domain-specific testing with CSR students and yielded positive results.

Introduction

This demo paper presents the design, development and evaluation of a novel visual analytics dashboard based on Social Set Analysis, Social Set Visualizer (SoSeVi), for big data analytics in the enterprise application domain of Corporate Social Responsibility (CSR). The presented dashboard connects social activity from Facebook with a thorough event timeline of the factory disasters in the Bangladesh garment industry. By using SoSeVi, even novice users can gain profound understanding of the industrial tragedies in Bangladesh, their background, and the resulting social media impact. Furthermore, the linked social media activity from eleven international companies in the garment industry can be interactively explored through different visualizations depicting actor mobility, conversation content, language distribution, and overall activity levels. The goal of this paper is to present a brief understanding of the design and development processes in implementing SoSeVi based on freely available open-source components in a robust, extensible manner. Moreover, an evaluation of SoSeVi is performed based on a task-based user study in conjunction with software and database performance optimization. The viability of the dashboard was assessed with special regard to the ease of use without prior training.

The remainder of the paper is organized as below. First, we briefly report on the garment industry sector in Bangladesh with regard to the enterprise application domain of CSR. Second, we outline the seven steps of our research methodology. Third, we present the technical description of the IT artifact, Social Set Visualizer (SoSeVi) including key features, technologies used, visualization framework, software and database optimization and an annotated screenshot. Fourth, we present some of the findings generated by using SoSeVi by researchers and practitioners. Fifth and last, we offer some conclusions and outline for future work.

Enterprise Domain: Global Supply Chains

The garment industry in Bangladesh is the second-largest exporter of clothing after China, and employs more than 3 million - mainly female - workers. The garment industry in Bangladesh has rapidly grown during the past 20 years while approving of lax safety regulations and frequent accidents [2]. On April 24th, 2013, factory disasters in the Bangladeshi garment sector culminated in the largest textile industry tragedy to date with the collapse of Rana Plaza, a factory building in an industrial suburb of Bangladesh's capital Dhaka in which more than 1100 garment workers died during the factory's collapse and fires (http://www.nytimes.com/2013/04/25/world/asia/bangladesh-building-collapse.html). The events of 2013 were reported by media outlets all over the world and deeply shocked many end-consumers of clothing products originating from Bangladesh. In various research publications, safety and struggles of workers in the Bangladesh garment industry have been widely discussed [3], also with special regard to ongoing protests [4], globalization-related problems [5] and ethical aspects of the factory disasters [6]. The factory disasters in Bangladesh prompted major fast fashion and discount clothing retailers like H&M and Walmart to join campaigns supporting textile workers' rights in Bangladesh. A more sustainable, but lagging impact is felt by the introduction of better methods of supply chain management such as social contracts in supply chains [7]. Our dashboard is a computational artefact that is situated in this particular nexus of social science research and enterprise practice.

Research Methodology

Our research methodology consisted of seven steps. First, we assembled a precise timeline of real-world events with respect to the Bangladesh factory accidents. Second, we generated a list of the traditional news media (print newspapers, TV and radio) reports of the real-world factory accidents in Bangladesh. Third, we reviewed the media reports and extracted a list of 11 multi-national companies that have been frequently mentioned in the traditional media reports

in relation to the Bangladesh garment factory accidents. Fourth, since Strategic Corporate Social Responsibility communication is conducted by companies on their Facebook pages, we extracted the full archive of the social data from the company walls of the 11 brands using SODATO [8]. Fifth, we designed, developed and evaluated the Social Set Visualizer dashboard of this Facebook corpus of approximately 180 million data points. Sixth, we addressed and answered a set of research questions using the dashboard. Seventh and last, we deployed the dashboard (with access control) to support ongoing research by CSR researchers and practitioners.

IT-Artefact: Social Set Visualizer (SoSeVi)

Data Collection & Processing

Event timeline of Bangladesh factory accidents and media reports were collected by desk research including systematic searches on the web and media databases Facebook data was collected from the Social Data Analytics Tool (SODATO) [8-10]. SODATO-provided Facebook activity datasets are generated as independent files for each company's Facebook wall, and were combined into one for using them as a whole data set that can be filtered or expanded on demand. Figure 1 SoSeVi's system schematic for the data acquisition, processing and visualization. The general concept follows the stages of the "Big Data Value Chain" introduced by Miller and Mork [11], with steps of preparation, organization and integration of the data prior to visualization and analysis. Data preparation tasks are performed in a pre-processing step which converts all CSV files to from their character encoding *UTF-16* to the more commonly used *UTF-8* and handles edge cases in which the generated SODATO output lacks proper data type encapsulation. Subsequently, a data normalization phase performs sanity checks on the input data and identifies malformed data or unneeded information. Lastly, all distinct data sets are aggregated while conserving information regarding their original source in an additional variable. The aggregated data is then imported into a database management system (DBMS), from which it can be accessed for visual analytics purposes.

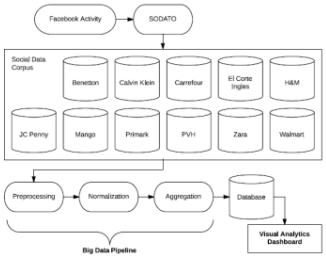


Figure 1. SoSeVi System Architecture

SoSeVi: Software Architecture

The dashboard consists of a client-side and a server-side part. Both need to be implemented with focus on client- and server-side performance in mind. This is depicted by HCI-informed ease of use and traditional scaling.

The server-side web application is based on the Node.js programming which provides server-side Javascript. In contrast, the server-side application is mainly based on the Express.js framework that provides a basic HTTP server. HTML templates are created using the Jade templating language. All client-side components of the SoSeVi dashboard are implemented in Javascript that runs within the end users' web browser. The dashboard is reinforced by several programming frameworks such as Backbone.js which is used for view management. Furthermore, several convenience frameworks such as underscore.js (utility functions) and moment.js (date handling) are used. All user-facing visualizations shown within the client-side Visual Analytics dashboard are implemented using dynamic SVG graphics which are built upon the powerful D3.js visualization framework.

SoSeVi: Visualization Framework

The technology choice for realizing the dashboard visualizations is the D3.js Javascript-based visualization framework which uses dynamic SVG images for data visualization. D3.js constitutes a lightweight and very extendable Javascript visualization framework. It uses cutting-edge web technology in order to facilitate the creation of complex visualizations. The flexibility provided by D3.js enables the creation of new kinds of interactive visualizations which are able to run on any device with decent processing resources. This includes Windows, MacOS and Linux based systems with screen sizes up to 4K devices. Figure 2 presents the software architecture of SoSeVi.

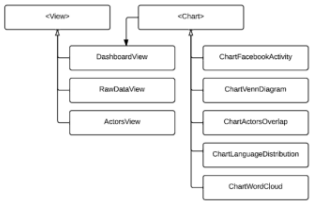


Figure 2. SoSeVi Software Architecture

DashboardView is the main view of the web application which contains the SoSeVi and is initially shown to the user. *RawdataView* presents a detailed search interface for the Facebook activity data. Many visualizations in DashboardView refer to RawDataView in order to provide the user with further information. *ActorsView* presents a dedicated interface for analysis tasks related to Actor Mobility across time and space of companies' Facebook walls. The visualizations of actor mobility in DashboardView refer to ActorsView in order to provide the user with further details when requested. Furthermore, ActorsView presents a handy set of tools for analysis of actor mobility and cross-postings between different time frames and Facebook walls.

SoSeVi: Dashboard Interface

Figure 3 presents the SoSeVi dashboard and its constituent visualizations for the full dataset.

The *Facebook Activity* visualization displays the social media activity on Facebook over the whole time period. It consists of a large main chart and a smaller mini chart underneath. Both charts use a line plot to display activity. The mini-chart reveals the full distribution of historical activity

and can also be used as a brush to dynamically isolate and a time period of the data shown in the main chart for detailed analysis. The *Actor Mobility across Space* visualization at the top right of the dashboard displays the number of different Facebook walls on which Actors have posted. For this visualization, a bar chart is used for rapid comparison. The chart depicts the number of Actors based on the number of Facebook walls they have posted to.

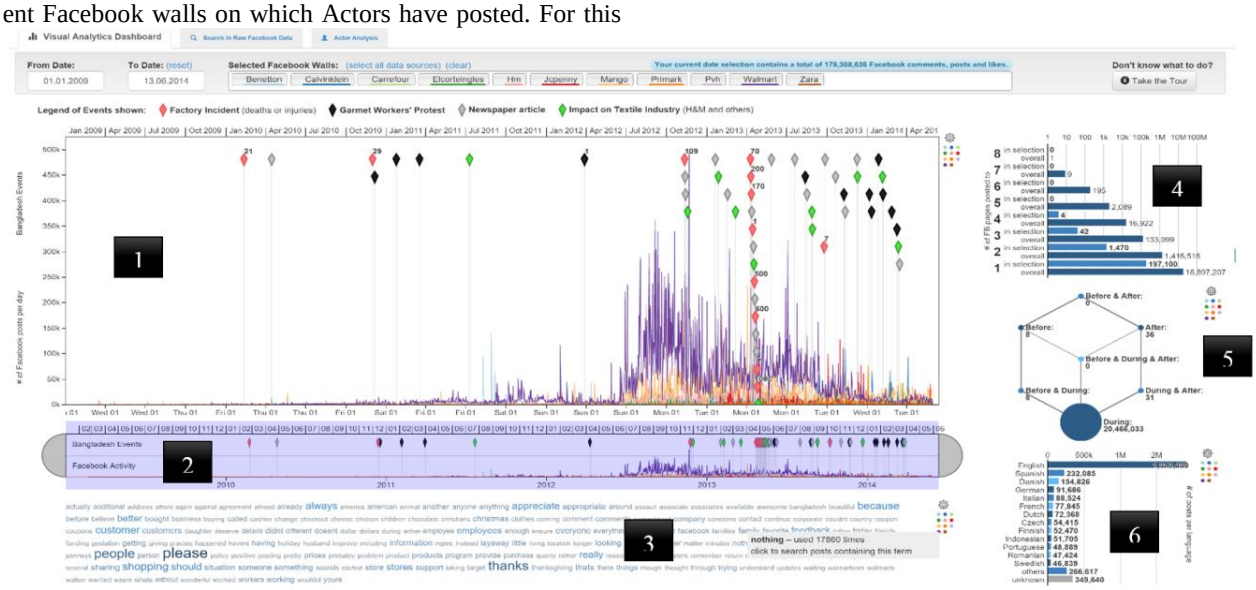


Figure 3: Social Set Visualizer: For the selected time period (see date range fields in top-left) and selected facebook walls (see colour coded selection bubble chart next to visualizations) → [1] Facebook Activity Chart; [2] Timeline of Bangladesh Factory Accidents & Facebook Actions; [3] Word Cloud of Text from Posts and Comments; [4] Actor Mobility across Space (facebook walls); [5] Actor Mobility across Time (before, during, after time-period of selection and combinations of them); and [6] Language Distribution

The *Actor Mobility across Time* visualization at the center right of the dashboard displays the number of Actors within each time period and their respective overlaps. For this visualization, an exploded Venn diagram is used which is aligned hexagonally to more clearly reveal cross sections than a traditional Venn diagram.

The *Language Distribution* visualization at the bottom right of the dashboard displays the number of social media Artifacts based on their language, again utilizing a bar chart for the most clear and effective comparisons. It presents each language and the respective number of social media activities during the selected timeframe.

The *Word Cloud* visualization located right beneath the Facebook Activity chart displays the results of the word frequency analysis based on conversation artifacts in the available social data. The font size of each word is determined by its overall frequency within all conversations that happened during the selected time period. The word cloud may not reveal the precise numerical differences that a table would, but it offers the most intuitive understanding of word frequency in a compact use of space.

Legend for Event Timeline is placed at the very top of the dashboard between the user-driven filtering interface and the Facebook Activity visualization. It conveys information about different types of events that are part of the event timeline. In the case at hand, the event timeline is based on the Bangladesh factory disaster events, which means that the event types classified are encoded in the legend.

The *user-driven filtering* interface contains two components. On the left hand side, the user may input Start and End Date of the timeframe to be visualized in the dashboard. Mouse or touch interactions with the input fields will reveal a hidden date picker component based on Boostrap Datepicker. This date picker enables the user to either input dates using a keyboard or specifying the day, month and year using their mouse or even a touch screen. Secondly, on the right hand side, the user may select the companies whose Facebook walls are shown in the Visual Analytics dashboard. User interaction with the available input field can be performed in various ways. The user can directly type Facebook walls into the field, which are then displayed in the visualization. An alternative method is that the user selects an item from a dropdown menu that appears when the input field is focused.

To summarize, the Big Data Visual Analysis dashboard empowers users to use it in different ways. The dashboard adheres to the user's preferred interaction method without making any assumptions. This means tablet users may also type in their selection of the Facebook walls, or desktop users may use the Datepicker to manually select a date.

Key Questions and Selected Findings

The following were some of the key questions investigated by using the SoSeVi:

- What is the distribution of the keyword "bangladesh" across time and space of 11 different Facebook walls?
- What is the period of maximum density of Bangladesh factory accidents?
- What, if any, are the trends and patterns with the word cloud chart, actor mobility across time chart (before, during, and after), actor mobility across space chart (Facebook walls), and language distribution during maximum accident density time period?
- In terms of social actors, who are the most influential negative critics of all 11 companies during the maximum accident density time period?
- In terms of social actors, who are the most influential positive loyalists of all 11 companies during the maximum accident density time period?
- In terms of social actors, what were the different strategies adopted by the Facebook wall admins of the 11 companies during maximum accident density time period?
- In terms of social actors, what were the different kinds of direct appeals to the CEOs of the companies made during maximum accident density time-period?

Due to space restrictions, we can't report all the findings from the visual analytics case study. That said, selected findings are that (a) the global supply chain concerns with regard to Bangladesh garment factories have been expressed by Facebook users from as far back as 2009, (b) there are many instances of authentic displays of support and expressions of empathy from Facebook users as well as robotic incidents of 'slactivism', (c) many of the uses of the word "please" were in relation to opening requests for new stores in the case of H&M, (d) protestors and activists employed different social media strategies on the different Facebook walls of companies but with little evidence for social influence (in terms of the number of likes and comments on their posts), (e) similarly, companies followed not only different CSR strategies but also different social media strategies before, during and after the Bangladesh garment factory accidents, (f) for almost all of the accidents, a majority of the users posting during the news-cycle (that is, traditional media coverage of the event) don't return to the Facebook walls again. That is, social media engagement during factory accidents is detected as episodic and 'bursty' with little overlap to the "business-as-usual" period before or after the accident.

Conclusion and Outlook

Computational social science research has reached a point where social media activity is ubiquitous yet hard to collect and analyze in domain-specific ways (with the notable exception of epidemiology). In conjunction with complex event timelines as depicted by the Bangladesh garment factory disasters, the data at hand presents numerous opportunities for attaining deep insights. In this context, visual analytics present the means of reaching those insights to many users with different backgrounds, both experts and novices alike. The novel implementation of the present Social Set Visualizer (SoSeVi) dashboard showcases that the creation of visual analytics software, which meets the high technical, analytical and user experience requirements of present-day computing, is viable (and can be achieved by an academic research group with limited resources). Furthermore, the developed IT artefact leverages open-source visual analytics frameworks to maximum extent in order to achieve a pure implementation of important concepts in visual analytics.

References

[1] Vatrapu, R., Mukkamala, R., and Hussain, A.: 'A Set Theoretical Approach to Big Social Data Analytics: Concepts, Methods, Tools, and Findings', in Editor (Ed.)^(Eds.): 'Book A Set Theoretical Approach to Big Social Data Analytics: Concepts, Methods, Tools, and Findings' (2014, edn.), pp. 22-24

[2] Sato, H.: 'Cournot Competition and Reduction of Corruption to Prevent Garment Factory Fires in Bangladesh', Advances in Management and Applied Economics 2014, 4, (4), pp. 17-20

[3] Khanna, P.: 'Making labour voices heard during an industrial crisis: workers' struggles in the Bangladesh garment industry', TRAVAIL, capital et société, 2011, 44, (2)

[4] Himi, S.A., and Rahman, A.: 'Workers Unrest in Garment Industries in Bangladesh: An Exploratory Study', Journal of Organization and Human Behaviour, 2013, 2, (3), pp. 49-55

[5] Rahman, S.: 'Broken Promises of Globalization: The Case of the Bangladesh Garment Industry' (Lexington Books, 2013. 2013)

[6] Stewart, K.L.: 'An ethical analysis of the high cost of low-priced clothing', proceedings of Northeastern Association of Business, Economics, and Technology (NABET), 2013, pp. 128

[7] Islam, M.A., Deegan, C., and Gray, R.: 'Social Audits and Multinational Company Supply Chain: A Study of Rituals of Social Audits in the Bangladesh Garment Industry', Available at SSRN 2466129, 2014

[8] Hussain, A., and Vatrapu, R.: 'Social Data Analytics Tool (SODATO)', in Tremblay, M., VanderMeer, D., Rothenberger, M., Gupta, A., and Yoon, V. (Eds.): 'Advancing the Impact of Design Science: Moving from Theory to Practice' (Springer International Publishing, 2014), pp. 368-372

[9] Hussain, A., and Vatrapu, R.: 'Social Data Analytics Tool: Design, Development and Demonstrative Case Studies', Proceedings of IEEE 18th International Enterprise Distributed Object Computing Conference (EDOC 2014), Ulm, Germany, 2014, pp. 414-417, ISBN: 978-411-4799-5467-4794/4714, DOI: 4710.1109/EDOCW.2014.4770

[10] Hussain, A., Vatrapu, R., Hardt, D., and Jaffari, Z.: 'Social Data Analytics Tool: A Demonstrative Case Study of Methodology and Software', in Cantijoch, M., Gibson, R., and Ward, S. (Eds.): 'Analyzing Social Media Data and Web Networks' (Palgrave Macmillan, 2014), pp. 99-118

[11] Miller, H.G., and Mork, P.: 'From data to decisions: a value chain for big data', IT Professional, 2013, 15, (1), pp. 57-59

Publication III

Social Set Visualizer II: Interactive Social Set Analysis of Big Data

---

Benjamin Flesch, Raghava Rao Mukkamala, Abid Hussain and Ravi Vatrapu. *Social Set Visualizer (SoSeVi) II: Interactive Social Set Analysis of Big Data.* In SetVR@ Diagrams, pages 19–28, 2016

Social Set Visualizer (SoSeVi) II: Interactive Social Set Analysis of Big Data

Benjamin Flesch[1], Ravi Vatraapu[1,2], Raghava Rao Mukkamala[1], and Abid Hussain[1]

[1] Centre for Business Data Analytics (http://bda.cbs.dk),
Department of IT Management, Copenhagen Business School, Denmark
[2] Westerdals Oslo School of Arts, Comm & Tech, Norway
{bf.itm, rv.itm, rrm.itm, ah.itm}@cbs.dk

Abstract Current state-of-the-art in big social data analytics is largely limited to graph theoretical approaches such as social network analysis (SNA) informed by the social philosophical approach of relational sociology. This paper proposes and illustrates an alternate holistic approach to big social data analytics, social set analysis (SSA), which is based on the sociology of associations, mathematics of set theory, and advanced visual analytics of event studies. We first presented the SSA approach to a wider audience at IEEE Big Data 2015 [6], IEEE EDOCW 2015 [7], and IEEE EDOCW 2016 [8]. Since then we worked on improving SoSeVi in order to further demonstrate its usefulness in large-scale visual analytics tasks of individual and collective behavior of actors in social networks. The current iteration of the Social Set Visualizer (SoSeVi) builds upon some of the concepts laid out by the *UpSet* project [14] and aims to further improve the capabilities of researchers and practitioners in big social data analytics alike. We then illustrate our new approach by reporting on the design, development, and evaluation results of a state-of-the-art visual analytics dashboard, the Social Set Visualizer (SoSeVi). The development of the dashboard involved cutting-edge open source visual analytics libraries (D3.js) and creation of new visualizations such as visualizations of actor mobility across time and space, conversational comets, and more. Evaluation of the dashboard consisted of technical testing, usability testing, and domain-specific testing with CSR students and yielded positive results. In conclusion, we discuss the new analytical approach of social set analysis and conclude with a discussion of the benefits of set theoretical approaches based on the social philosophical approach of associational sociology.

Keywords: Big Social Data, Social Set Analysis, Computational Set Analysis, Big Data Visual Analytics, Event Studies

1 Introduction

This paper introduces a new research approach situated in the domains of Data Science [4,15,22] and Computational Social Science [13] with practical applications to Big Social Data Analytics in organizations [26,24,23]. It addresses one of the important theoretical and methodological limitations in the emerging paradigm of Big Data Analytics of social media data [25]. In particular, it address the major limitation in existing research on Big Social Data analytics that computational methods, formal models and software tools are largely limited to graph theoretical approaches [9] (such as SNA [2]), and are informed by the social philosophical approach of relational sociology [5]. There are no other unified modeling approaches to social data that integrate the conceptual, formal, software, analytical and empirical realms [20]. This results in a research problem when analyzing Big Social Data from platforms like Facebook and Twitter, as such data consists of not only dyadic relations but also individual associations [21]. For Big Social Data analytics of Facebook or Twitter data, the fundamental assumption of SNA, that social reality is constituted by dyadic relations and interactions that are determined by structural positions of individuals in social networks [19], is neither necessary nor sufficient [27].

For example, consider a Facebook post made on the official Facebook wall of Lionel Messi, the soccer prodigy who plays for FC Barcelona and Argentina's national football team. Each official post by Messi to his Facebook page typically receives more than 100,000 likes, 25,000 comments and 18,000 shares. Such association-based and content-driven social media interactions involving large number of social actors are unlike the other social interactions such as face-to-face, email, phone and instant messaging in the sense that what binds the interacting social actors together in the first instance is not so much the relational ties (strong vs. weak ties) but associations ranging from the player himself, the teams that he plays for, to the cultural, ethnic, national and linguistic attributes. Modeling such Facebook interactions using affiliation networks creates the problem of an extremely low number of nodes with an extremely high number of nodes as spokes. Further, such SNA assumes the central social psychological concept of "homophily" that social actors with similar interests (that is, associations) prefer to interact with each other. To overcome this limitation and address the research problem, this paper proposes an alternative holistic approach to Big Social Data analytics that is based on the sociology of associations as well as the mathematics of set theory and offers to develop fundamentally new methods and tools for Big Social Data analytics, Social Set Analysis (SSA). Our overarching research question is stated as: *How, and in what way, can methods and tools for Social Set Analysis, derived from the alternative holistic approach to Big Social Data analytics based on the sociology of associations and the mathematics of set theory, result in meaningful facts, actionable insights and valuable outcomes?*

The rest of the paper is organized as follows. First, we present a philosophical template for holistic approaches to computational social sciences, compare and contrast the dominant approach of social network analysis with the proposed novel approach of social set analysis and discuss the benefits of set theoretical

approaches based on the social philosophical approach of associational sociology in section 2. Then, we illustrate our new analytical approach by reporting on the design and development of our state-of-the-art visual analytics dashboard, the Social Set Visualizer (SoSeVi), that builds on and extends the UpSet visualizations of set intersections.

2 Theoretical Framework

The theoretical concepts behind our proposed approach of Social Set Analysis are discussed here.

2.1 Set Theoretical Big Social Data Analytics

Social Set Analysis (SSA) as employed in this paper is concerned with the mobility of social actors across time and space. For mobility across time, we conduct SSA of big social data from the Facebook walls of eleven companies from the same industry with an analytical focus on the set of actors that interacted with the company before, during and after the real-world events, and set theoretical intersections of the three time periods. Similarly, for mobility across space, we conduct set inclusions and exclusion of actors who interacted with different Facebook walls. This will allow us to uncover not only the interactional dynamics over time and space but also identify actor sets that correspond to marketing segmentations such as brand loyalists, brand advocates, brand critics and social activists.

2.2 Event Study Methodology

Event studies is a finance methodology to assess an impact on corporate wealth (e.g. stock prices) caused by events such as restructuring of companies, leadership change, mergers & acquisitions [3,17,16]. It has been a powerful tool since the late 1960s to assess financial impact of changes in corporate policies and used exclusively in the area of investments and accounting to examine stock price performance and the dissemination of new information [1].

While there is no unique structure for event study methodology, at a higher level of abstraction, it contains identifying three important time periods or windows. First, defining an event of interest and identifying the period over which it is active (event window), the second involves identifying the estimation period for the event (pre-event or estimation window) and the final one being identifying the post-event window [16]. In social set analysis of a real-world event, we have applied event study methodology to identify the three important time periods of user interactions on social media platforms: *before* (pre-event window), *during* (event window) and *after* (post-event window).

3 Related Work

The improved version of the Social Set Visualizer (SoSeVi) has been influenced by the previous work of the *UpSet* project [14], and the visualizations of set intersection based on innovative approaches to combination matrices. In contrast to UpSet, SoSeVi uses server-side calculations on its underlying big social data corpus, and therefore is able to handle much larger volumes of data with 100,000s to millions of actors moving between set intersections. Both projects strive to provide a real-time visual analytics tool.

The previous version of SoSeVi used Venn diagrams to showcase actor migration between time periods and different Facebook walls , with focus shifting towards the use of Euler visualizations as discussed in [18] for version 2 of SoSeVi.

4 The Visual Analytics Tool

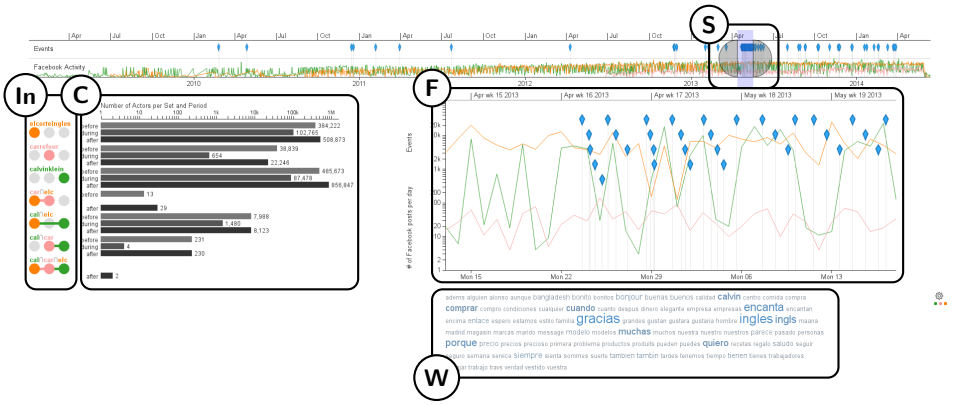


Figure 1. Social Set Visualizer showing 8M Facebook entries from the *Carrefour*, *CalvinKlein* and *El Corte Ingles* Facebook pages adapted to the computational set analysis approach: [F] main activity chart zoomed in on the user-selected time period using the [S] selection tool. Underneath the dynamically calculated word cloud [W] is located. On the left side we see all set intersections [In] displayed as a combination matrix, where [C] displays the cardinality of each individual set intersection over the *Before*, *During* and *After* periods. On clicking or hovering, we see a visualization of period-over-period actor migration between set intersections.

We showcase version II of the Social Set Visualizer (SoSeVi) tool which is adapted to the challenges of visualizing user-selected and dynamically calculated large-scale set intersections depicting aggregated actor behavior in social networks such as Facebook. Figure 1 illustrates the major design shift from the SoSeVi version which was presented in [6].

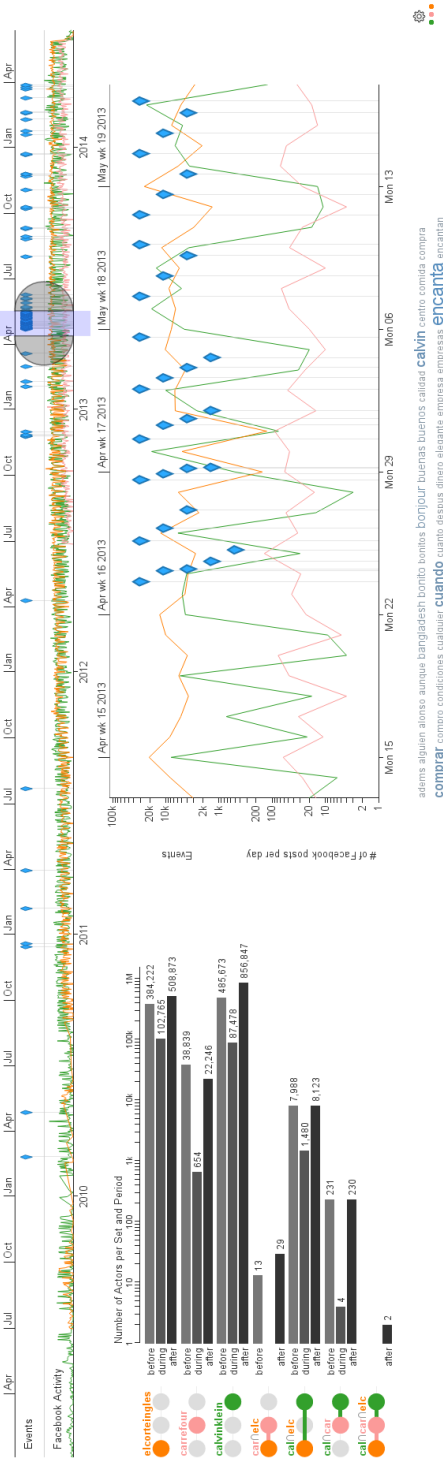


Figure 2. Social Set Visualizer II with interactive selection of time period for analysis and calculation of intersections for *Before*, *During* and *After* periods over the dataset fetched from Facebook. Set intersections are dynamically encoded through a combination matrix. The migration of actors between time periods and set intersections is showcased through cardinality changes on the left hand side. User-provided event markers signify important real-world events with relevance for the underlying research domain of the investigating analyst.

Figure 1 depicts the *DashboardView* which is the central interface of the web application. It contains all visualizations and is initially shown to the user. It consists of both small and large overall social media activity visualizations [F]. The researcher can use the time period selection tool [S] to navigate the data, and to toggle data sets from different Facebook walls depending on the analysis tasks at hand. Based on the user-selected time period, which we label as the *During* period, the tool is able to deduct *Before* and *After* time periods by looking at the beginning (earliest event) and end (most recent event) of the underlying data. An alphabetical word cloud [W] underneath the main activity chart [F] illustrates the most important conversation topics in the during period. This showcases the pluggable architecture of SoSeVi, which facilitates diverse real-time content analysis tasks on the underlying data, all based on a user-selected time frame for analysis. To the left of the main activity visualization [F], set intersections [In] are dynamically visualized based on the user selection of the time period. Set intersections are encoded in a combination matrix. Each data source uses a distinct color.

For each set intersection, we render up to three bar graphs in [C]. One bar each is drawn for every single *Before*, *During* and *After* period, when the underlying set has a cardinality of at least one actor. The bars are horizontally stacked, with the topmost bar signifying the *Before* period, the center bar *During*, and the lowest *After*. More information on the visualization of actor migration through time and space sets is shown in figure 3.

When clicking on or hovering over the set cardinality visualizations in [C], a visualization of actor migration between periods and set intersections is displayed as shown in figure 4. SoSeVi calculates all possible set intersection for each set with all sets of the following period (migrations from *Before* to *During*, and migrations from *During* to *After*) based on the user-selected time period using the selection tool [S]. Cardinality numbers illustrate the actual migration volume, whereas the right-hand side bar labels indicate the destination of each migration.

Augmenting the extensive visual analytics features of SoSeVi, *RawdataView* presents a detailed search interface for the underlying Facebook activity data. It is accessible to the user through various means by interacting with the visualizations of the DashboardView. ActorsView presents a dedicated interface for analysis tasks related to Actor Mobility across time and space of companies' Facebook walls. The visualizations of actor mobility in DashboardView refer to ActorsView in order to provide the user with further details when requested. ActorsView presents a handy set of tools for analysis of actor mobility and cross-postings between different time periods and Facebook walls.

The Social Set Visualizer can be accessed at http://bigdata:bigdata@5.9.5.20/ with user name and password *bigdata*. It has been tested by using Webkit and Gecko-based web browsers.

4.1 Data Acquisition

Facebook data was collected through the Social Data Analytics Tool (SODATO) [11,10,12]. SODATO-provided Facebook activity datasets are generated as independent files

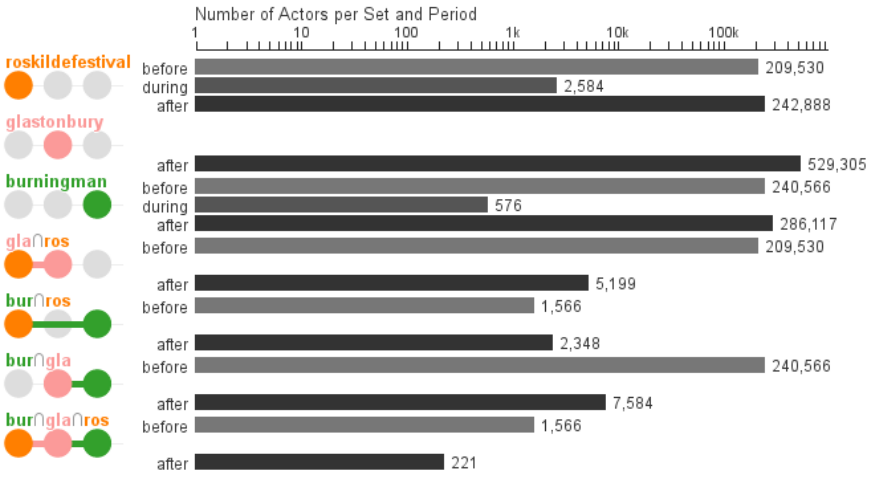


Figure 3. Visualization of set intersections and set intersection cardinality *Before*, *During*, and *After* the user-selected time period, illustrating the distribution of social media actors over time and space.

Figure 4. Visualization of actor migration concerning the *Calvin Klein* Facebook wall *During* the user-selected time period, showcasing strength and destinations of migrations. Migrations towards the Calvin Klein Facebook wall from the *Before* to the *During* Period are displayed in red color. Migrations originating from the Calvin Klein *During* period are received by to other intersections' *After* periods, and displayed in green color. The cardinality of each migration movement concerning the user-selected period and set is clearly visible in red (incoming migration) or green color (outgoing migration).

for each company's Facebook wall, and were combined into one for using them as a whole data set that can be filtered or expanded on demand.

5 Software Development

The SoSeVi dashboard is implemented as a client-side web application, and uses the D3.js Javascript SVG visualization framework. D3.js constitutes a lightweight and very extendable Javascript visualization framework which can display visualizations for a multitude of browser-based clients. The flexibility provided by D3.js enables the creation of new kinds of interactive visualizations which are able to run on any device with decent processing resources including Windows, MacOS and Linux based systems with screen sizes up to 4K, which gives SoSeVi the flexibility needed for various purposes in different research areas.

After thorough (re-)evaluation of Apache Spark and other NoSQL-based storage solutions, the decision to use PostgreSQL for data storage was not overthrown due to lack of empirically measured benefits in execution time with our test data. In version 2 of SoSeVi, all set intersection calculations have been outsourced from PostgreSQL to Redis. A dedicated Redis instance now performs memory-intensive set intersection calculations with a significantly better execution speed and pipes the calculation results back to the user-facing dashboard in real time.

6 Conclusion

Version 2 of the Social Set Visualizer (SoSeVi) project presented in this paper provides a significantly better interface for Social Set Analysis (SSA) tasks of big social data originating from Facebook. We showcase our interactive tool for large-scale real-time set intersection calculations to the research community. SoSeVi 2 depicts the first visual analytics tool to visualize migration flows between set intersections in big social data.

7 Future Work

We strive to add more customization features in order to provide the user with more viable investigation strategies, such as sorting and filtering of sets. This was also demonstrated in the *UpSet* project, but real-time implementation of those features was out of scope for version 2 of SoSevi. Add extension points to perform statistical calculations over the set intersections and improve the overall measurement of migration flows. The research tool will be extended for non-Facebook data social media data.

8 Acknowledgments

The authors were partially supported by the project Big Social Data Analytics: Branding Algorithms, Predictive Models, and Dashboards funded by Industriens

Fond (The Danish Industry Foundation). Any opinions, findings, interpretations, conclusions or recommendations expressed in this paper are those of its authors and do not represent the views of the Industriens Fond (The Danish Industry Foundation).

References

1. Binder, J.: The event study methodology since 1969. Review of quantitative Finance and Accounting 11(2), 111–137 (1998)
2. Borgatti, S.P., Mehra, A., Brass, D.J., Labianca, G.: Network analysis in the social sciences. Science 323(5916), 892–895 (2009)
3. Bromiley, P., Govekar, M., Marcus, A.: On using event-study methodology in strategic management research. Technovation 8(1), 25–42 (1988)
4. Cleveland, W.S.: Data science: an action plan for expanding the technical areas of the field of statistics. International Statistical Review 69(1), 21–26 (2001), http://dx.doi.org/10.1111/j.1751-5823.2001.tb00477.x
5. Emirbayer, M.: Manifesto for a relational sociology. The American Journal of Sociology 103(2), 281–317 (1997)
6. Flesch, B., Vatrapu, R., Mukkamala, R.R., Hussain, A.: Social set visualizer: A set theoretical approach to big social data analytics of real-world events. In: Big Data (Big Data), 2015 IEEE International Conference on. pp. 2418–2427 (Oct 2015)
7. Flesch, B., Hussain, A., Vatrapu, R.: Social set visualizer: Demonstration of methodology and software. In: 2015 IEEE 19th International Enterprise Distributed Object Computing Workshops and Demonstrations (EDOCW). pp. 148–151 (Sept 2015)
8. Flesch, B., Vatrapu, R.: Social set visualizer (sosevi) ii: Interactive computational set analysis of big social data. In: 2016 IEEE 20th International Enterprise Distributed Object Computing Workshops and Demonstrations (EDOCW) (in press/2016)
9. Gross, J.L., Yellen, J.: Graph theory and its applications. CRC press (2005)
10. Hussain, A., Vatrapu, R.: Social data analytics tool: Design, development, and demonstrative case studies. In: Enterprise Distributed Object Computing Conference Workshops and Demonstrations (EDOCW), 2014 IEEE 18th International. pp. 414–417 (Sept 2014)
11. Hussain, A., Vatrapu, R.: Social data analytics tool (sodato). In: DESRIST 2014. Lecture Notes in Computer Science (LNCS). Springer, vol. 8463, pp. 368–372 (2014)
12. Hussain, A., Vatrapu, R., Hardt, D., Jaffari, Z.: Social data analytics tool: A demonstrative case study of methodology and software. In: Analysing Social Media Data and Web Networks. Palgrave Macmillan (2014)
13. Lazer, D., Pentland, A., Adamic, L., Aral, S., BarabĞsi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., Van Alstyne, M.: Computational social science. Science 323(5915), 721–723 (2009)
14. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., Pfister, H.: Upset: Visualization of intersecting sets. IEEE Transactions on Visualization and Computer Graphics (IEEE InfoVis '14) (2014), live Demo: http://vcg.github.io/upset
15. Loukides, M.: What Is Data Science? O'Reilly Media (2012)
16. MacKinlay, A.C.: Event studies in economics and finance. Journal of economic literature pp. 13–39 (1997)

17. McWilliams, A., Siegel, D.: Event studies in management research: Theoretical and empirical issues. Academy of management journal 40(3), 626–657 (1997)
18. Micallef, L., Rodgers, P.: euler ape: Drawing area-proportional 3-venn diagrams using ellipses. PloS one 9(7), e101717 (2014)
19. Mizruchi, M.S.: Social network analysis: Recent achievements and current controversies. Acta sociologica 37(4), 329–343 (1994)
20. Mukkamala, R.R., Hussain, A., Vatrapu, R.: Towards a formal model of social data. IT University Technical Report Series TR-2013-169, IT University of Copenhagen, Denmark (November 2013)
21. Mukkamala, R.R., Hussain, A., Vatrapu, R.: Towards a set theoretical approach to big data analytics. In: 3rd International Congress on Big Data (IEEE BigData 2014) (June 2014)
22. Ohsumi, N.: From data analysis to data science. In: Data Analysis, Classification, and Related Methods, pp. 329–334. Springer Berlin Heidelberg (2000), http://dx.doi.org/10.1007/978-3-642-59789-3_52
23. Sponder, M.: Social media analytics: effective tools for building, intrepreting, and using metrics. McGraw-Hill (2012)
24. Sterne, J.: Social media metrics: How to measure and optimize your marketing investment. John Wiley & Sons (2010)
25. Tufekci, Z.: Big questions for social media big data: Representativeness, validity and other methodological pitfalls. arXiv preprint arXiv:1403.7400 (2014)
26. Vatrapu, R.: Understanding social business. In: Emerging Dimensions of Technology Management, pp. 147–158. Springer (2013)
27. Vatrapu, R., Mukkamala, R.R., Hussain, A., Flesch, B.: Social set analysis: A set theoretical approach to big data analytics. In: IEEE Access, 4. pp. 2542–2571 (2016)

All links were last followed on July 14th, 2016.

Publication IV

A Big Social Media Data Study of the 2017 German Federal Election based on Social Set Analysis of Political Party Facebook Pages with SoSeVi

---

Benjamin Flesch, Ravi Vatrapu and Raghava Rao Mukkamala. *A Big Social Media Data Study of the 2017 German Federal Election Based on Social Set Analysis of Political Party Facebook Pages with SoSeVi*. In Big Data (Big Data), 2017 IEEE International Conference on, pages 2720–2729. IEEE, 2017

A Big Social Media Data Study of the 2017 German Federal Election based on Social Set Analysis of Political Party Facebook Pages with SoSeVi

Benjamin Flesch[1], Ravi Vatrapu[1,2], Raghava Rao Mukkamala[1]
[1]Copenhagen Business School, Denmark and [2]Westerdals Oslo School of Arts, Comm & Tech, Norway
{bf.digi, rv.digi, rrm.digi}@cbs.dk

***Abstract*—We present a big social media data study that comprises of 1 million individuals who interact with Facebook pages of the seven major political parties *CDU*, *CSU*, *SPD*, *FDP*, *Greens*, *Die Linke* and *AfD* during the 2017 German federal election. Our study uses the Social Set Analysis (SSA) approach, which is based on the sociology of associations, mathematics of set theory, and advanced visual analytics of event studies. We illustrate the capabilities of SSA through the most recent version of our Social Set Analysis (SoSeVi) tool, which enables us to deep dive into Facebook activity concerning the election. We explore a significant gender-based difference between female and male interactions with political party Facebook pages. Furthermore, we perform a multi-faceted analysis of social media interactions using gender detection, user segmentation and retention analysis, and visualize our findings. In conclusion, we discuss the analytical approach of social set analysis and conclude with a discussion of the benefits of set theoretical approaches based on the social philosophical approach of associational sociology.**

***Keywords*—Big social media data, Social set analysis, Big data visual analytics, Facebook, 2017 German federal election, Bundestagswahl, CDU, CSU, SPD, FDP, Grüne, AfD, Linke**

I. Introduction

This paper applies *Social Set Analysis* research approach to the 2017 federal election in Germany, more precisely to the activity on the major political parties' Facebook walls. *Social Set Analysis* is a research approach situated in the domains of Data Science [1]–[3] and Computational Social Science [4] with practical applications to Big Social Data Analytics in organizations [5]–[7]. It addresses one of the important theoretical and methodological limitations in the emerging paradigm of Big Data Analytics of social media data [8]. In particular, it address the major limitation in existing research on Big Social Data analytics that computational methods, formal models and software tools are largely limited to graph theoretical approaches [9] (such as SNA [10]), and are informed by the social philosophical approach of relational sociology [11]. There are no other unified modeling approaches to social data that integrate the conceptual, formal, software, analytical and empirical realms [12]. This results in a research problem when analyzing Big Social Data from platforms like Facebook and Twitter as such data consists of not only dyadic relations but also individual associations [13]. For Big Social Data analytics of Facebook or Twitter data, the fundamental assumption of SNA that social reality is constituted by dyadic relations and interactions that are determined by structural positions of individuals in social networks [14] is neither necessary nor sufficient [15]. Previous versions of the *Social Set Visualizer* tool have been introduced to showcase the *Social Set Analysis* approach [16].

For example, consider a Facebook post made on the official Facebook wall of Lionel Messi, the soccer prodigy who plays for FC Barcelona and Argentina's national football team. Each official post by Messi to his Facebook page typically receives more than 100,000 likes, 25,000 comments and 18,000 shares. Such association-based and content-driven social media interactions involving large number of social actors are unlike the other social interactions such as face-to-face, email, phone and instant messaging in the sense that what binds the interacting social actors together in the first instance is not so much the relational ties (strong vs. weak ties) but associations ranging from the player himself, the teams that he plays for, to the cultural, ethnic, national and linguistic attributes. Modeling such Facebook interactions using affiliation networks creates the problem of an extremely low number of nodes with an extremely high number of nodes as spokes. Further, such SNA assumes the central social psychological concept of "homophily" that social actors with similar interests (that is, associations) prefer to interact with each other. To overcome this limitation and address the research problem, this paper proposes an alternative holistic approach to Big Social Data analytics that is based on the sociology of associations and the mathematics of set theory and offers to develop fundamentally new methods and tools for Big Social Data analytics, Social Set Analysis (SSA). Our overarching research question is stated as, *How, and in what way, can methods and tools for Social Set Analysis derived from the alternative holistic approach to Big Social Data analytics based on the sociology of associations and the mathematics of set theory result in meaningful facts, actionable insights and valuable outcomes?*

The rest of the paper is organized as follows. First, we present a philosophical template for holistic approaches to computational social sciences, compare and contrast the dominant approach of social network analysis with the proposed novel approach of social set analysis and discuss the benefits of set theoretical approaches based on the social philosophical approach of associational sociology in Sec II. Second, we present the most recent version of our Social Set Visualizer (SoSeVi) tool in III.

Third, we take a deep dive into Facebook activity concerning the 2017 German federal election held on 24th of September 2017 on a political party level. Section IV illustrates

the capabilities of SoSeVi by showcasing growth and retention of audience by political parties, user segmentation into loyalists and persons with positive and negative feelings towards a political party, and further analyses based on first names and gender classification.

Fourth and last, we discuss the findings from our illustrative case study, offer methodological and analytical reflections on social set analysis, identify its limitations, and outline future work directions. We have not provided any dedicated section for related work, but we have referred the relevant literature at appropriate places throughout the paper.

II. Theoretical Framework

Social Set Analysis (SSA) as employed in this paper is concerned with the mobility of social actors across time and space. For mobility across time, we conduct SSA of big social data from the Facebook walls of the seven major political parties in Germany, with an analytical focus on the set of actors that interacted with the parties during the 2017 federal election campaign. Similarly, for mobility across space, we conduct set inclusions and exclusion of actors who interacted with different Facebook walls. This will allow us to uncover not only the interactional dynamics over time and space but also identify actor sets that correspond to marketing segmentations such as loyalists, advocates, critics and activists. The theoretical framework and the formal model behind our proposed approach of Social Set Analysis have been elaborated in previous papers such as [16] [15].

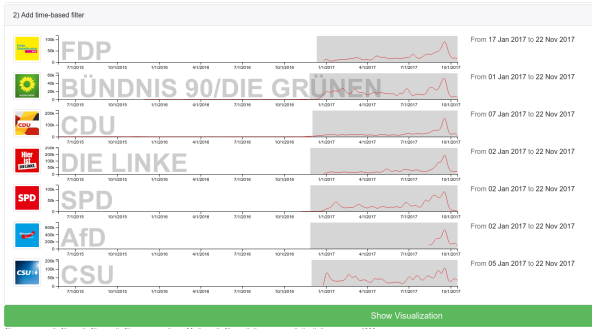
III. Social Set Visualizer (SoSeVi) tool

A. User interface

The Social Set Visualizer (SoSeVi) tool for Social Set Analysis has been under active development since 2014, with older version presented to the research community in several papers. The latest version focuses on Upset-inspired [17] visualization set intersections, and is paired with a built-in Facebook crawler. The set intersection visualization allows researchers to define social media interactions in a set query language, and then perform further analysis based on the set of individuals at hand which resulted from the query.

Figure 1(a) showcases the latest version of the Social Set Analysis user interface provided by the Social Set Visualizer tool. After selection of Facebook pages of interest, the user can compare these Facebook pages in an Upset-inspired [17] set visualization tailored to the Social Set Analysis approach. Social Set Visualizer (SoSeVi) provides means to segment individuals on social media and visualize their interactions. Word cloud visualization and aggregated Facebook page information is shown in figure 1(b).

To summarize, the SoSeVi big data visual analysis dashboard empowers users to use it in many different ways. The dashboard adheres to the user's preferred interaction method without making any assumptions. This means tablet users may also type in their selection of the Facebook walls, or desktop users may use the date picker to manually select a date. The dashboard may be accessed at http://rf2017.roonk.de/.



(a) Selection of Facebook pages and time period of interest as preparation for a visualization of set overlaps and intersection cardinalities.

(b) Facebook page overview with alphabetical word cloud and Facebook reaction visualization.

Figure 1: User interface provided by Social Set Visualizer.

B. Technology

The technology choice for realizing the dashboard visualizations is the D3.js Javascript-based visualization framework which uses dynamic SVG images for data visualization. D3.js constitutes a lightweight and very extendable Javascript visualization framework which can display visualizations for a multitude of browser-based clients. The flexibility provided by D3.js enables the creation of new kinds of interactive visualizations which are able to run on any device with decent processing resources including Windows, MacOS and Linux based systems with screen sizes up to 4K devices. Data is stored in a relational database and heavily indexed using *PostgreSQL*. Queries are cached both in database tables and in-memory using *Redis*.

IV. 2017 German federal election case study

A. Background

The 2017 German federal election held on 24th of September 2017 was the largest political event in recent years. Major topics such as the European migrant crisis [18], central bank policies [19] and workplace equality [20] have put pressure on incumbent Angela Merkel, her cabinet and the political parties *CDU*, *CSU* and *SPD* closely affiliated with her. Both pro-business liberal party *FDP* and the green party *Bündnis '90 / Die Grünen* aim to get more foothold with mainstream voters than in previous years.

More extreme political parties such as recently formed *Alternative für Deutschland* (AfD) and leftist party *Die Linke* contest voters' mind share and aim to get more influence in the future government. Based on these circumstances, we take a deep dive into social media reactions on the major political parties' Facebook pages to better understand the state of mind of the political parties' audiences and ultimately, the German voters.

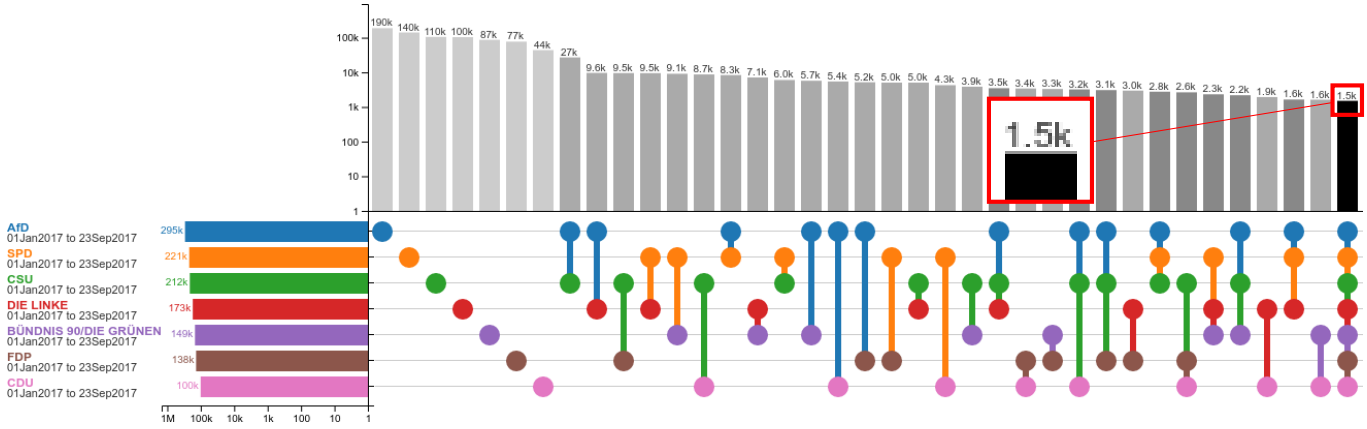


Figure 2: SoSeVi-based set visualization of Facebook audience overlap between major political parties in Germany during the 2017 federal election campaign. The overlap between all seven political parties is represented by the black bar on the right side of the visualization. The grid-based set overlap visualization using interconnected circles is inspired by Upset [17].

B. Methodology

Our research methodology consisted of several steps. First, we fetched the Facebook walls of the major political parties in Germany: Angela Merkel's *CDU*, Bavarian *CSU*, social democratic *SPD*, liberal *FDP*, green party *Bündnis '90 Die Grünen*, leftist party *Die Linke* and ultra-conservative alternative party *AfD*. For this, we use a self-made Facebook crawler. Furthermore we restrict our observation timeframe to the beginning of 2017 up until the day before the federal election, 23rd of September 2017. Second, we analyze collected Facebook activity with our *Social Set Visualizer* (SoSeVi) tool. We visualize overlaps between individual parties' Facebook audiences and illustrate inner-party retention rates throughout the hot phase of the election campaign. Third, we perform deep dives into audience segments of interest and illustrate the capabilities of SoSeVi by addressing party loyalist, audience reactions and demographics such as the most common first names of individuals interacting with the Facebook walls. Fourth, we discuss our findings and deploy the dashboard internally to support ongoing research.

C. Data Collection & Processing

The event timeline of the 2017 federal election was collected through desk research including systematic searches in web and media databases. Facebook data was previously collected through the Social Data Analytics Tool (SODATO) [21]–[23]. For this paper a SoSeVi-internal crawler was used to provide Facebook data shown in table I. The general concept follows the stages of the "Big Data Value Chain" introduced by Miller and Mork [24], with steps of preparation, organization and integration of the data prior to visualization and analysis. The aggregated data is then imported into a database management system (DBMS), from which it can be accessed for visual analytics purposes.

D. Size of political party Facebook audience

In figure 2 SoSeVi is utilized to visualize a total of 958,834 individuals who interacted with German political party Facebook pages during the 2017 federal election. This number is also displayed in figure IV as all-party total. We examine the aggregate number of individuals that interacted with each parties' Facebook page during the examination period up to 23rd of September 2017, as visualized through the left-side horizontal bar chart in figure 2.

Party	Posts	P.Reactions	Comments	C.Reactions
AfD	970	2,107,255	445,978	1,031,180
CDU	550	374,830	152,904	364,261
CSU	598	985,812	142,078	455,527
FDP	652	592,527	80,403	106,132
GREEN	442	361,351	97,309	214,113
LINKE	609	607,137	104,082	246,823
SPD	531	719,632	121,215	229,401

Table I: Overview of Facebook dataset of major German political parties

It strikes that newcomer *AfD* leads with a total of 295,000 individuals, followed in second place by social democrats *SPD* who interacted with 221,000 individuals. Third largest is Bavarian-only *CSU* party with 212,000 individuals active on their page, the sister party of Angela Merkel's *CDU*. *CDU* themselves are in last place, because only 100,000 individuals interacted with their Facebook page during the 2017 federal election campaign. All minor parties such as the *FDP* with 138,000 individuals, the *Green party* with 149,000 and the leftist party *Die Linke* with 173,000 had Facebook interactions with more unique individuals than Angela Merkel's ruling party *CDU*.

E. Audience overlap between political party Facebook pages

In figure 2 we also visualize overlaps of Facebook audiences between the major political parties in Germany in the 2017 federal election period from 1st of January to 23rd of September 2017. We use Social Set Analysis approach to

	April	May		June		July		August		September		All months	
Party	#	% chg	#	% chg	#	% chg	#	% chg	#	% chg	#	Sparkline	CMGR
AfD	63.1k	-8%	58.4k	8%	63.4k	-6%	60.0k	35%	92.8k	40%	155.0k		19.7%
CDU	11.8k	16%	14.1k	18%	17.2k	32%	25.4k	-11%	22.8k	51%	46.7k		31.7%
CSU	39.9k	-18%	33.7k	11%	37.9k	33%	56.8k	-22%	46.5k	30%	66.0k		10.6%
FDP	25.1k	6%	26.8k	-25%	21.5k	34%	32.5k	29%	45.8k	25%	61.1k		19.5%
GREEN	18.6k	11%	20.9k	17%	25.1k	-47%	17.1k	58%	40.4k	10%	44.9k		19.3%
LINKE	18.7k	52%	39.3k	0%	39.3k	-2%	38.5k	8%	41.8k	52%	86.7k		35.9%
SPD	31.4k	-18%	26.7k	27%	36.8k	-10%	33.5k	54%	72.5k	13%	83.8k		21.7%

Table II: Monthly growth rate of unique individuals who interacted with German political party Facebook pages during the 2017 federal election campaign between 1st of January and 23rd of September 2017. Sparklines visualize month with lowest and highest number of individuals on Facebook page. Compound monthly growth rate is calculated and compared.

calculate sets of individuals and visualize overlaps between the sets at hand. Major two-set overlaps between political parties are:

1) We observe that more than 27,000 individuals were active both on the *CSU* and the *AfD* Facebook pages, displaying the biggest audience overlap between two political parties.
2) The second major audience overlap is between *AfD* and leftist party *Die LINKE* with 9,600 individuals.
3) The third largest overlap is between Bavarian *CSU* party and liberal *FDP* party with more than 9,500 individuals active on both parties' Facebook pages.
4) Fourth largest overlap is between social democrats *SPD* and leftist *Die Linke* with *9,500* individuals, followed by fifth largest overlap between *SPD* and the *Green* party with 9,100 individuals active on both Facebook pages.
5) Angela Merkel's *CDU* and her Bavarian sister party *CSU* depict the sixth largest overlap with 8,700 individuals.

Further overlaps between political party Facebook audiences are visualized in the figure, but due to space restrictions we cannot list all of them. The major overlaps identified seem to follow the parties' closeness on the political spectrum, even though at the moment we cannot explain the detailed reason for the relative differences in cardinality between overlaps such as *CSU*/*AFD* and *SPD*/*Die Linke*.

F. Audience growth during election campaign

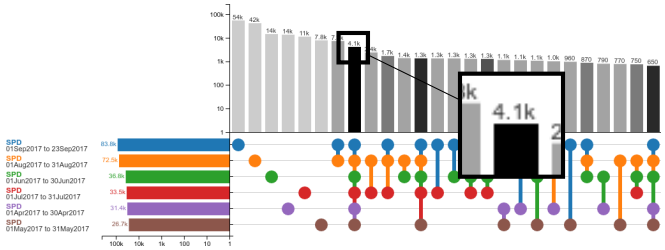
The audience growth rate in terms of the total number of individuals who were active on a certain political party's Facebook page during the campaign is showcased in table II. Using social set analysis we create sets of individuals who interacted with a certain party for each month of the election campaign. Cardinalities of monthly sets for each political party have been taken from the set visualizations of figure 3. Based on this data, a compound monthly growth rate (CMGR) has been calculated to compare each party's audience growth during the time period of the election campaign. We observe the following:

1) For all parties, the final month of campaigning, September, was the best month in terms of total number of individuals they interacted with.
2) No party showcases a steady, consistent growth story. All of them have at least one month where they actually decreased their audience compared to the previous month.
3) Comparing the compound monthly growth rate (CMGR), both leftist *LINKE* (+35.9%) and Angela Merkel's party *CDU* (+31.7%) depict the biggest growth over the whole period of investigation. Both are also the only parties where both April is the overall weakest month and September the overall peak.
4) With only 10.6% over the whole campaign, *CSU* showcased the lowest overall growth rate.
5) All other parties *SPD*, *FDP*, *GREEN*, and *AfD* expressed a compound growth rate of around 20% per month.
6) In August, penultimate month of the 2107 election campaign, current chancellor Angela Merkel's parties *CDU* and *CSU* both decreased in the number of individuals that interacted with their Facebook pages by a total of 69.3k people (-11% and -22% respectively). This is interesting because one would expect that during August, at the peak of campaigning, both sister parties would continue to push very hard. This decrease could be explained with summer holidays for the shared campaigning team.
7) Also in August, *SPD*, the biggest rival of *CDU*/*CSU*, grew their audience at 54%. With a total of 72.5k individuals, *SPD* reached a larger audience on Facebook than both *CDU* (22.8k) and CSU (46.5k) combined.

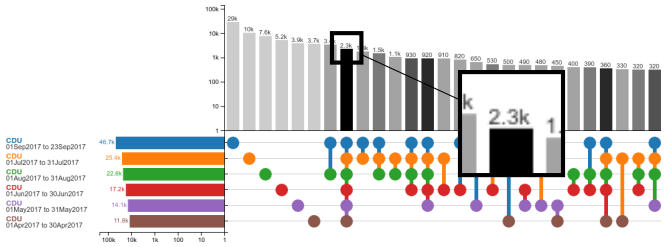
G. Audience retention of political party Facebook pages

We visualize month-over-month retention of Facebook audience for each political party in Germany from 1st of April up to election day 24th of September 2017. For this purpose we create six monthly slices (April, May, June, July, August, September until 23rd) for each party and utilize SoSeVi to perform social set analysis on them.

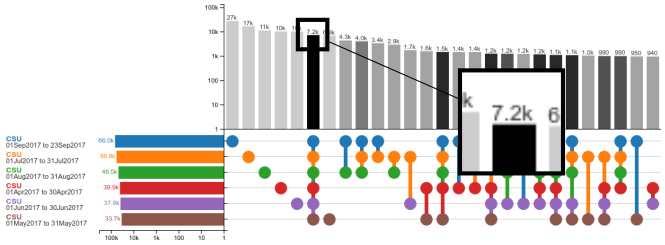
Using the example of social democrat party *SPD*, we visualize in Figure 3(a) the month-to-month development of individuals interacting with the party's Facebook page. The visualization shows that 4,100 individuals interact with the party on a monthly level, and the vast majority of users interact with the party's Facebook page on a very loose basis. Even though we see a steady month-to-month growth between April and September, the retention of individuals seems to be lacking. In September up until election day, a total of 83,000 individuals interacted with *SPD*, but 54,000 of those only did so in September and not in any prior months. The visualization for social democrat *SPD* party can be accessed



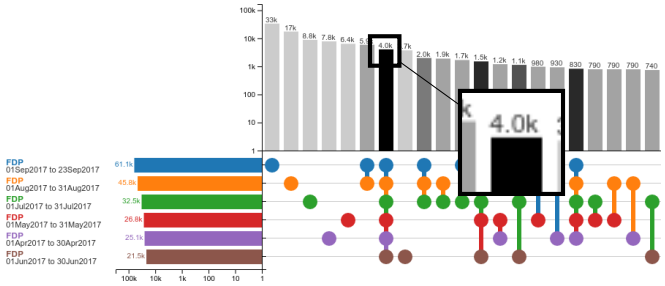
(a) SPD



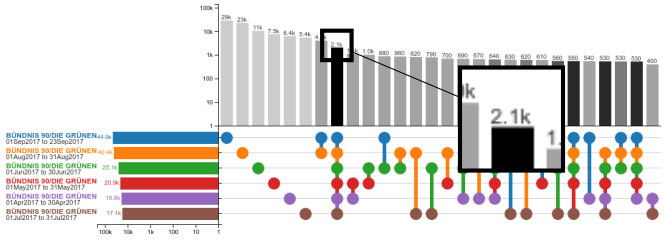
(b) CDU



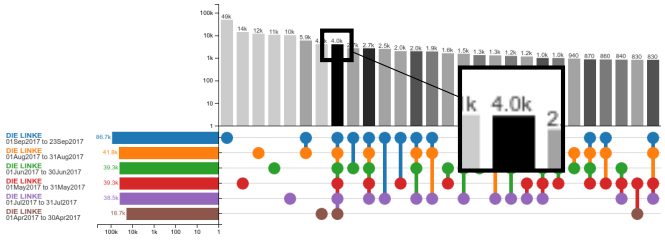
(c) CSU



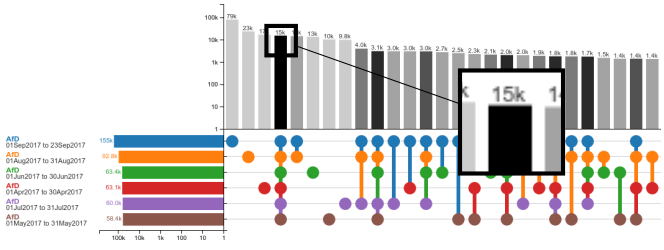
(d) FDP



(e) Green party



(f) Leftist party



(g) AfD

Figure 3: SoSeVi-based visualization of month-over-month development and retention of Facebook audience for German political parties, sliced monthly until election day. Loyalist audience for each party is depicted by the black vertical bar spanning all six month-based sets.

online at rf2017.roonk.de/upset. Likewise we visualize audience retention for other political parties in figure 3.

H. Identification of political party loyalists on Facebook

We define political party loyalists as the set of individuals who interact with a certain party's Facebook page at least once per month. For this purpose we examine monthly slices for the six months preceding election day, same as in previous section IV-G. We determine loyalist audience from the number of individuals who are active on a specific party's Facebook page in every single month within the observation period up to election day. Using figures 3(a), 3(b), 3(c), 3(d), 3(e), 3(f), 3(g) we can determine the total number of loyalists for each political party.

In order to put the absolute size of party loyalist audience in perspective, we compare official party membership numbers to the size of the loyalist audience on Facebook and calculate a ratio. Number memberships has been collected for each party from official publications [25]. Table III showcases that loyalist Facebook audience varies highly between political parties. The massive membership bases of the two major parties *SPD* and *CDU* are not significantly more active on the parties' Facebook pages than the loyalist audiences of smaller parties.

Relative to the total number of party memberships, small parties such as *AfD*, *FDP*, and the Leftist party *Die Linke*

Party	Total party members	# loyalists on Facebook	% of members
SPD	432,706	4,100	0.95%
CDU	431,920	2,300	0.53%
CSU	142,412	7,200	5.06%
GREEN	61,596	2,100	3.41%
LINKE	58,910	4,000	6.79%
FDP	53,896	4,000	7.42%
AfD	26,409	15,000	56.80%

Table III: Comparison of political party loyalist audience on Facebook and official party membership numbers. Membership numbers are from 31. December 2016 and based on official publications [25].

Party	M/F ratio party memberships*	All individuals						Reaction to Posts			Reaction to Comments			Writing Comments		
		Total	Female	Male	N/A	%	M/F Ratio	Female	Male	M/F Ratio	Female	Male	M/F Ratio	Female	Male	M/F Ratio
AfD	**5.25**	**294,951**	77,514	188,409	29,028	9.84%	**2.43**	56,768	146,909	**2.59**	34,613	70,435	**2.03**	22,011	60,539	**2.75**
CDU	**2.85**	**100,494**	25,371	62,839	12,284	12.22%	**2.48**	17,016	40,635	**2.39**	9,575	22,879	**2.39**	6,021	19,757	**3.28**
CSU	**4.00**	**212,019**	58,728	136,509	16,782	7.92%	**2.32**	50,216	115,146	**2.29**	17,494	36,853	**2.11**	10,149	29,149	**2.87**
FDP	**3.35**	**137,717**	31,414	96,032	10,271	7.46%	**3.06**	26,732	80,585	**3.01**	6,778	21,322	**3.15**	5,095	20,875	**4.10**
GREEN	**1.56**	**148,626**	59,325	72,102	17,199	11.57%	**1.22**	51,037	48,682	**0.95**	11,934	24,215	**2.03**	6,625	18,450	**2.78**
LINKE	**1.70**	**172,902**	50,006	100,937	21,959	12.70%	**2.02**	42,904	80,663	**1.88**	14,925	30,792	**2.06**	7,639	22,540	**2.95**
SPD	**2.13**	**220,904**	71,674	119,264	29,966	13.57%	**1.66**	60,460	90,220	**1.49**	15,237	31,983	**2.10**	9,370	27,908	**2.98**
All parties	**2.53**	**958,834**	296,772	548,827	113,235	11.81%	**1.85**	254,271	461,346	**1.81**	88,916	176,913	**1.99**	55,121	149,985	**2.72**

Table IV: First name based gender classification of social media actors on political party Facebook pages during the 2017 German federal election. Official party member gender ratio is based on 2016 data published by German federal ministry for political education (BPB) [26]. *N/A* displays failed gender classification.

interact with a high number of individuals compared to their total memberships. *AfD* in particular is rapidly growing with a low number of official party memberships, thus the high percentage of 56.80%. Compared with peers, the *Green* party receives only a small amount of loyalist interaction on the Facebook page, both in absolute numbers but also as a relative percentage to their peers in terms of party memberships *Die Linke* and *FDP*.

I. Audience reactions to political party Facebook posts

Table V showcases Facebook reactions by individuals to posts by political parties. For this analysis we count the number of individuals who interact with the party post with a Facebook reaction, focusing on the most widely used Facebook reactions LOVE, LIKE, SAD, ANGRY and HAHA. We observe:

1) Far-right *AfD* received reactions from more than 225k audience members. This is 40k more people than the next biggest parties, *CSU* (180k) and *SPD* (175k).
2) Every party except Angela Merkel's *CDU* received reactions from more than 110k individuals to their Facebook postings. In total, only 66k users reacted to *CDU* posts.
3) Used by more than 90% of all individuals, *LIKE* depicts major audience reaction to political party posts.
4) Liberal *FDP* receives a *LIKE* from 95% of their interacting Facebook audience.
5) Receiving *LIKEs* from 202k individuals, far-right *AfD* significantly eclipses Angela Merkel's *CDU* which only receives *LIKE*s from 56k individuals during the campaign.
6) Far-right *AfD* received *ANGRY* reactions from 51k individuals or 23% of their audience.
7) Reactions other than *LIKE* are not very frequently used, major exception being the numerous *ANGRY* reactions towards *AfD* posts.

Party	# individuals with reaction to posts by a political party										# unique individuals
	ANGRY		SAD		HAHA		LOVE		LIKE		
CSU	16,751	9%	8,389	5%	20,858	12%	5,952	3%	165,783	92%	**179,348**
AfD	51,179	23%	11,854	5%	37,084	16%	21,144	9%	201,968	90%	**225,160**
SPD	5,332	3%	5,399	3%	12,676	7%	11,689	7%	159,372	91%	**175,649**
LINKE	8,933	6%	4,044	3%	6,039	4%	12,528	9%	133,882	94%	**142,540**
GREEN	5,867	5%	4,656	4%	8,961	8%	9,152	8%	100,961	89%	**113,722**
FDP	3,893	3%	2,556	2%	6,910	6%	5,940	5%	110,054	95%	**115,658**
CDU	4,093	6%	2,386	4%	7,811	12%	5,077	8%	56,363	84%	**66,718**

Table V: Audience reactions to political party Facebook posts during 2017 election campaign.

J. Comparing Facebook gender distribution with official party membership data

Table IV displays the results of gender-based Facebook audience segmentation. In the first part of the table, we show aggregate numbers for each political party. The center of the table shows audience reactions to posts and to comments by the political party, aggregated by gender. Furthermore, we show male/female comment authorship. In the final row of the table we plot the official male to female ratio based on party membership publications for comparison. We perform audience segmentation by gender to showcase the full potential of Social Set Analysis. Audience interactions with political party Facebook pages are analyzed along this dimension. Gender inference is performed based on the first name of the Facebook user at hand. We use the $nam\_dict.txt$ database[1] to link first names with genders. This technique for gender inference has been successfully applied by other researches such as [27]. Based on male/female audience segmentation of German political party Facebook walls as shown in table IV we can point out several qualitative findings:

1) Both on an aggregate and on an individual level, discussion and reactions on the political parties' Facebook pages appear male-dominated, with a male-to-female ratio as high as 4.10 for comment authorship on *FDP* page.
2) The only exception to this observation are reactions to posts on the *Green party*'s page. With post reactions from 51,037 females and only 48,682 males, this is the only dimension in table IV where we can count more females than males interacting with the party's posts.
3) Incumbent ruling party *CDU* has the fewest individuals on their Facebook page, less than half as many as their biggest rival, the social democrat party *SPD*.
4) Leftist party *Die Linke* is the only political party where the Male/Female ratio of all dimensions of interaction (post and comment reactions, comment authorship) with their Facebook page is higher than the official Male/Female ratio based on their party memberships.
5) Apart from the Leftist party *Die Linke*, all other parties have a more balanced male-to-female ratio on their Facebook page than the male-to-female ratio based on official party memberships numbers suggests.

[1] nam_dict.txt first-name based gender classification database (c) 2008 Jörg Michael, available at https://www.heise.de/ct/ftp/07/17/182/

(a) Female first name distribution across political parties

First name / Party	PETRA	SABINE	ANDREA	ANNA	CLAUDIA	JULIA	SANDRA	MONIKA	NICOLE	KARIN	SUSANNE	SARAH	HEIKE	LISA	MARIA	ANJA	MARTINA	KERSTIN	BIRGIT	MARION	CHRISTINE	TANJA	KATHARINA	LAURA	BRIGITTE	MELANIE	DANIELA	ELKE	BARBARA	ANNE	EVA	ANGELIKA	NADINE	MANUELA	STEFANIE	GABRIELE	CHRISTINA	SIMONE	ALEXANDRA	UTE	RENATE	JANA	SABRINA	SONJA	MICHAELA	LENA	NINA	KATJA	INGRID	KATRIN
AfD	.34	.31	.30	.21	.26	.18	.30	.25	.27	.21	.18	.18	.23	.17	.18	.21	.19	.21	.17	.20	.17	.19	.11	.10	.14	.18	.17	.16	.12	.11	.12	.16	.16	.19	.14	.14	.12	.14	.12	.13	.11	.12	.15	.13	.13	.08	.08	.12	.11	.12
CDU	.29	.31	.28	.24	.26	.21	.19	.23	.22	.22	.22	.15	.19	.14	.22	.17	.19	.17	.17	.17	.18	.14	.17	.14	.15	.13	.13	.15	.16	.18	.13	.18	.12	.13	.12	.15	.12	.13	.11	.13	.11	.11	.07	.10	.12	.10	.07	.09	.11	.11
CSU	.47	.35	.37	.20	.33	.17	.24	.41	.21	.34	.23	.08	.24	.13	.24	.17	.23	.21	.26	.24	.25	.15	.13	.07	.29	.13	.17	.22	.23	.12	.17	.26	.10	.20	.14	.21	.14	.15	.15	.16	.23	.09	.10	.15	.18	.05	.07	.09	.21	.10
FDP	.26	.27	.26	.25	.25	.26	.20	.18	.20	.17	.19	.16	.14	.16	.15	.18	.15	.15	.16	.13	.13	.12	.18	.14	.12	.11	.14	.12	.13	.15	.15	.11	.13	.09	.12	.13	.12	.10	.13	.11	.09	.10	.07	.08	.08	.12	.13	.10	.11	.11
GREEN	.41	.47	.40	.56	.36	.48	.32	.30	.28	.28	.35	.38	.26	.38	.29	.27	.24	.24	.25	.20	.27	.19	.30	.34	.20	.24	.20	.19	.23	.29	.27	.16	.19	.13	.20	.18	.19	.18	.16	.17	.13	.19	.15	.21	.15	.27	.23	.19	.13	.18
LINKE	.33	.30	.28	.29	.25	.25	.23	.23	.20	.22	.20	.22	.20	.23	.18	.19	.19	.20	.18	.18	.16	.14	.16	.16	.15	.13	.14	.16	.15	.18	.15	.15	.14	.12	.12	.13	.11	.12	.13	.13	.13	.14	.10	.11	.10	.12	.15	.13	.11	.12
SPD	.34	.33	.29	.34	.30	.33	.27	.26	.26	.22	.22	.28	.25	.28	.22	.22	.20	.18	.19	.18	.17	.19	.20	.23	.18	.18	.16	.19	.16	.18	.16	.17	.19	.13	.15	.14	.15	.15	.16	.15	.13	.14	.17	.15	.14	.14	.12
All parties %	.35	.32	.31	.31	.29	.28	.27	.26	.26	.23	.23	.23	.22	.22	.22	.21	.20	.20	.20	.19	.19	.18	.18	.18	.18	.17	.17	.17	.17	.17	.17	.16	.16	.15	.15	.15	.15	.14	.14	.14	.14	.14	.14	.14	.13	.13	.13	.13	.13	.13
# Actors [1000s]	3.3	3.1	3.0	2.9	2.8	2.7	2.6	2.5	2.5	2.2	2.2	2.2	2.1	2.1	2.1	2.0	1.9	1.9	1.9	1.8	1.8	1.7	1.7	1.7	1.7	1.7	1.7	1.6	1.6	1.6	1.6	1.6	1.6	1.5	1.4	1.4	1.4	1.4	1.4	1.4	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.2
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50

(b) Male first name distribution across political parties

First name / Party	MICHAEL	THOMAS	CHRISTIAN	ANDREAS	PETER	DANIEL	STEFAN	FRANK	MARTIN	MARKUS	JAN	HANS	SEBASTIAN	ALEXANDER	MATTHIAS	KLAUS	SVEN	PATRICK	TOBIAS	MARCEL	FLORIAN	JÜRGEN	JENS	MARCO	WOLFGANG	UWE	RALF	DAVID	DIRK	OLIVER	BERND	PHILIPP	JÖRG	DENNIS	ROBERT	CHRISTOPH	MAX	ALEX	SASCHA	MARC	CHRIS	TIM	MARIO	KEVIN	FELIX	STEPHAN	LUKAS	JOHANNES	PAUL	STEFFEN
AfD	1.6	1.4	1.1	1.2	1.1	.93	.89	.96	.75	.73	.52	.62	.55	.57	.55	.59	.65	.60	.45	.61	.40	.57	.58	.56	.49	.59	.53	.46	.51	.43	.47	.29	.48	.46	.44	.30	.31	.38	.44	.36	.42	.29	.53	.43	.22	.33	.20	.21	.31	.39
CDU	1.6	1.4	1.1	1.2	.88	.81	.84	.79	.70	.71	.64	.62	.61	.62	.59	.52	.50	.46	.46	.46	.48	.45	.48	.43	.45	.41	.43	.37	.43	.46	.44	.42	.41	.39	.32	.48	.37	.30	.37	.35	.27	.37	.28	.28	.36	.38	.36	.38	.31	.29
CSU	1.8	1.6	1.2	1.4	1.2	.81	1.1	.91	.81	.91	.49	.89	.61	.67	.65	.78	.48	.48	.51	.37	.57	.73	.49	.46	.72	.61	.56	.30	.45	.45	.59	.34	.48	.30	.44	.44	.38	.34	.34	.34	.34	.28	.35	.23	.26	.39	.26	.39	.24	.32
FDP	1.6	1.4	1.5	1.2	.90	1.1	1.0	.82	.87	.83	1.0	.59	.89	.92	.74	.52	.61	.64	.74	.53	.69	.48	.57	.49	.50	.40	.45	.50	.48	.57	.47	.80	.43	.49	.45	.60	.62	.44	.40	.51	.47	.56	.26	.36	.60	.41	.54	.51	.34	.36
GREEN	1.1	1.0	.78	.81	.66	.64	.64	.53	.63	.50	.58	.39	.49	.47	.45	.40	.35	.36	.38	.26	.40	.35	.34	.28	.34	.33	.31	.39	.31	.33	.33	.34	.28	.24	.30	.34	.30	.30	.23	.27	.27	.28	.21	.17	.32	.25	.29	.31	.24	.22
LINKE	1.3	1.0	.88	.94	.83	.82	.74	.71	.67	.53	.62	.53	.58	.54	.46	.47	.43	.47	.41	.43	.41	.43	.42	.42	.41	.42	.40	.43	.36	.35	.36	.35	.35	.36	.35	.31	.38	.40	.34	.29	.36	.33	.30	.32	.36	.28	.33	.25	.34	.27
SPD	1.1	.95	.80	.76	.63	.69	.63	.57	.58	.52	.61	.43	.52	.44	.44	.43	.41	.40	.42	.41	.41	.36	.33	.33	.37	.33	.32	.38	.33	.33	.32	.37	.29	.32	.27	.31	.32	.27	.30	.29	.26	.34	.21	.33	.33	.24	.31	.26	.24	.21
All parties %	1.2	1.0	.90	.90	.77	.75	.73	.66	.63	.60	.58	.54	.53	.52	.48	.47	.46	.46	.43	.43	.43	.42	.42	.40	.40	.40	.39	.39	.38	.37	.37	.36	.36	.36	.35	.34	.34	.33	.33	.33	.33	.32	.31	.31	.30	.29	.29	.28	.27	.27
# Actors [1000s]	12	10	8.7	8.6	7.4	7.2	7.0	6.3	6.0	5.8	5.5	5.2	5.1	5.0	4.6	4.5	4.4	4.4	4.1	4.1	4.1	4.0	3.8	3.8	3.8	3.7	3.7	3.6	3.5	3.5	3.4	3.4	3.4	3.4	3.3	3.3	3.2	3.1	3.1	3.1	3.1	3.0	3.0	2.9	2.8	2.8	2.7	2.6	2.6
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50

Table VI: Comparative visualization TOP50 most frequent male and female first names across German political party Facebook pages. Gender detection is performed based on first name. Colored areas display each party's Facebook audience having a certain first name as a percentage of the whole dataset. The number of actors at the bottom of each table concerns the absolute number of individuals in our dataset who hold a certain first name.

K. Gender-based differences in Facebook interactions with German political parties

Furthermore we examine whether there is a statistically significant difference between male and female individuals in their interaction with German political party's Facebook pages during the period of the 2017 federal election. For this purpose we perform a chi-square test of gender-based differences in engagement with 6 degrees of freedom. The test shows a significant difference between males and females, with $p < 0.05$ and $\chi^2 = 17825.46$.

Potential limitations of this finding are the extent and veracity of our first name based gender classification approach. We have manually verified gender classification results for the top 100 most frequently used first names, yet the long-tail correctness of classification results has not been thoroughly examined. The name lists underlying our gender classification approach is targeted at German-speaking population and does not capture all names from other cultural backgrounds. Table IV depicts gender classification results. A total of 113,235 actors (12%) have not been successfully classified. To further test for gender-based differences in Facebook interaction, we should assume that all non-gender-classified first names are female, and repeat the chi-square test. Again it shows that the finding is significant with $p < 0.05$ and $\chi^2 = 20944.96$.

L. Frequency analysis of first names across political party Facebook audience

In table VI we visualize top 50 most frequently occurring first names across all individuals interacting with political party pages during the election campaign. More specifically, table V(a) depicts the frequency distribution of overall top 50 female first names and how often these first names are observed in each political party within the time period of the election campaign. Table V(b) provides the same information for all individuals that were classified as males based on their first names.

The visualization of top 50 female first names in table V(a) provides insight into party-specific distribution of first names. Facebook audience of the *GREEN* party exhibits above-average frequency of female first names, e.g. *ANNA* (0.56% vs. global average 0.31%) and *JULIA* (0.48% vs. global average 0.28%). *CSU* displays higher variance than the *GREEN* party: With 0.08% of global audience, names such as *SARAH* are significantly less frequent on the *CSU* page than it would be expected given the 0.23% overall average. Conversely, table V(b) depicts overall top 50 male first names from our dataset and their frequency across political parties. As shown in section IV-J, political party Facebook audience within the 2017 German federal election campaign is overwhelmingly male. Viewing the male first name visualization this becomes apparent through the fact that most frequencies are about two to three times higher than in table V(a). No significant trends are visible to the eye.

The top 50 first names returned by the gender-focused approach in this section largely mimic historic demographics of Germany, and thus don't provide significant findings apart from several outliers and slight trends between political parties.

Table VII: Top 50 most uniquely attributable first names for each political party in the 2017 German federal election. Numbers depict the percentage share of all individuals with a certain first name interacting with the respective party's Facebook page.

(a) AfD

First name / Party	RONNY	SYLVIO	JAN	RINGO	GARY	ENRICO	SILVIO	RICO	RENE	DENNY	STEVE	PIOTR	MAIK	ROCCO	COLIN	ROY	ROBBY	ONKEL	MIKE	TINO	RICCARDO	MARIO	BRIAN	ANDREW	HANSI	ANDY	INGOLF	JEFF	ANTHONY	ADOLF	JAMES	DANNY	THOR	FALKO	MANDY	MIRKO	TOMMY	HEIKO	MARKO	ROGER	TONY	KEN	ANDRZEJ	JIM	BOB	JOHN	PAVEL	EDDY	RONALD	RICKY
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
AfD	64	63	62	60	59	58	57	57	57	57	56	55	54	54	54	54	54	53	53	53	52	52	51	51	51	51	50	50	50	50	50	50	50	49	49	49	49	48	48	48	48	48	48	48	48	48	48	48	48	48
CDU	9	12	6	5	3	10	8	9	8	7	8	9	10	12	6	4	7	5	10	11	11	9	6	7	9	10	6	12	6	7	10	8	11	11	7	10	7	11	10	9	10	6	9	8	7	10	7	9	10	5
CSU	20	20	16	22	13	19	21	19	22	16	16	14	19	21	11	16	17	25	24	24	16	24	16	16	35	24	28	11	10	33	15	17	23	19	13	19	21	27	22	24	12	12	15	18	18	15	16	21	29	20
FDP	9	11	10	9	12	8	9	9	10	6	9	5	11	5	14	9	10	6	14	13	11	12	10	12	9	13	14	13	8	7	8	13	16	11	6	14	12	14	12	16	8	9	9	10	11	12	10	7	14	8
GREEN	8	10	10	14	8	9	8	10	7	8	9	7	8	12	15	11	13	12	9	10	8	10	9	8	8	9	9	7	11	12	5	10	13	8	12	11	10	12	12	12	9	10	11	11	18	11	9	6	10	17
LINKE	18	12	9	22	12	18	20	18	16	16	16	11	17	20	15	18	21	13	16	17	18	17	14	14	14	18	19	12	8	9	14	18	21	13	17	18	17	17	14	15	21	22	13	20	14	16	14	22	15	21
SPD	12	10	11	10	12	15	11	14	15	10	12	16	16	15	17	11	12	10	16	13	8	15	21	12	14	14	13	25	21	18	17	15	21	13	15	19	16	17	16	15	13	18	11	15	18	17	15	17	16	14
# Total Actors	1824	124	168	123	134	1305	668	792	2371	333	950	136	2456	208	154	337	162	125	3032	882	163	4229	276	221	238	1866	197	130	132	152	309	1052	155	245	728	1223	587	2945	982	478	487	153	102	171	168	1194	111	233	772	112
# Actors in AfD	838	56	84	52	67	550	285	331	996	158	422	64	984	81	63	148	65	54	1129	332	68	1555	112	95	87	684	71	50	58	56	129	399	50	98	303	428	218	974	353	165	193	59	42	63	60	442	45	86	259	40

M. Top 50 most uniquely attributable first names for each political party

Table VII showcases an alternative approach to providing a unique perspective on the Facebook audience of German political parties. For each party, we identify the top 50 first names that are most uniquely attributable to the party at hand. We calculate relative percentage share of all audience members with a certain first name and select the top 50 highest percentage first names for each party. First names with less than 100 individuals and party names are filtered out. For example with AfD in table VII(a), we can see in the leftmost column that 64% of all individuals with the first name *Ronny* interact with the *AfD* Facebook page, while only 9% of all *Ronnys* interact with *CDU* page. The total number of individuals named *Ronny* in our data set is 1834, of which 838 (64%) interact with *AfD* during the campaign. We further examine the most uniquely attributable first names for each party and describe our findings:

1) *AfD* VII(a): Most uniquely attributable names are "stereotypical" for the eastern part of Germany. Frequency distribution heavily skewed towards *AfD*.
2) *CDU* VII(b): Mainly Arabic first names, but overall very low level of uniqueness (percentages less than 40%), many shared with *SPD*.
3) *CSU* VII(c): Traditional German names, both male and female, with percentages between 40 and 50%.
4) *FDP* VII(d): German male first names.
5) *GREEN* VII(e): German female first names.
6) *LINKE* VII(f): First names with some Turkish background, most likely related to immigrant workers during the early days of German federal republic.
7) *SPD* VII(g): First-ranked *TC* means *Türkiye Cumhuriyeti* (Republic of Turkey), Turkish activists added TC in front of their name to signal their support of Turkey during a shit storm including *SPD*. Most unique names related to *LINKE*, but Arabic names in long tail as shown in VII(b).

V. Discussion

Due to space restrictions, we presented only a subset of the empirical findings resulting from the use of the Social Set Visualizer (SoSeVi) tool by researchers and practitioners in various fields such as Corporate Social Responsibility (CSR), Computational Social Sciences (CSS) and healthcare. These empirical findings demonstrate the analytical utility of our proposed set theoretical approach to big social data and our social set analysis implementation in the SoSeVi visual analytics dashboard.

A. Reflections on the IT-Artifact

Computational social science research has reached a point where social media activity is ubiquitous yet hard to collect and analyze in domain-specific ways (with the notable exception of epidemiology). In conjunction with complex event timelines as depicted by the 2017 German federal election, the data at hand presents numerous opportunities for attaining deep insights. In this context, visual analytics present the means of reaching those insights to many users with different backgrounds, both experts and novices alike. The novel implementation of the present Social Set Visualizer (SoSeVi) dashboard showcases that the creation of visual analytics software, which meets the high technical, analytical and user experience requirements of present-day computing, is viable (and can be achieved by an academic research group with limited resources). Furthermore, the developed IT artifact leverages open-source visual analytics frameworks to maximum extent in order to achieve a pure implementation of important concepts in visual analytics.

B. Reflections on the Set Theoretical Approach

The current paradigm in computational social science is dominated by a theoretical focus on relationships of actors and artifacts, and the mathematical modeling of those relationships as social networks based on graph theory.

This leads to the big social data triumvirate of relational sociology (as candidate social philosophy), graph theory (as candidate mathematical and formal model), and social network analysis (as candidate analytical framework). Our argument is not that relational sociology, graph theory, and social network analysis are invalid or ineffective. Social Network approaches have proven their analytical suitability and ability in diverse application domains ranging from epidemiology to organizational behavior. Instead, our argument is that other candidate sociological approaches, mathematical theories, and analysis techniques need to be explored to further advance the field of computational social science. After all, relational sociology is just one of the many competing and co-existing theories in sociology describing, explaining and predicting social phenomena; along with process, ethnomethodology, structuration, identity, structural functionalism, cognitive and cultural theories. Our paper's primary contribution to not only to offer an alternate holistic approach of social theory (associations),

(b) CDU

Party \ First name	محمد	MAHMOUD	ANAS	MUHAMMAD	MOHAMAD	AHMAD	KHALED	AHMED	ALWIN	MOHAMMED	JAMAL	HUSSEIN	MOHAMMAD	MECHTHILD	BENEDICT	SAID	LUDGER	ADNAN	OMAR	MOHAMED	HAMID	ALBRECHT	REZA	SAMIR	SAMUEL	SAMI	GERO	MIC	HEI	ANDREE	ARND	JORGE	CLAUDIUS	AZIZ	BENEDIKT	ABDUL	MAGNUS	SÖNKE	TIMON	HERIBERT	HASSAN	FERDI	IRA	ANSGAR	MARTEN	TAREK	EDDA	GEORGE	FRIEDEL	CARL
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
AfD	5	13	15	10	11	12	10	14	35	14	11	12	9	16	19	7	20	9	17	14	12	22	11	9	21	13	20	32	44	40	26	35	26	10	18	8	20	20	20	35	9	22	22	20	26	17	19	31	33	25
CDU	36	34	32	31	30	29	27	27	26	25	25	24	24	24	24	21	21	21	20	20	19	19	19	19	18	18	18	18	18	18	18	17	17	17	17	17	17	17	17	17	17	16	16	16	16	16	16	16	16	16
CSU	9	5	11	7	8	7	8	10	26	3	8	6	7	27	18	5	30	7	7	8	4	34	8	9	18	9	15	28	34	22	27	15	31	11	29	9	19	14	13	39	10	32	20	17	17	10	28	18	37	24
FDP	6	3	6	3	0	3	2	4	19	2	4	3	4	13	33	2	20	5	5	6	4	18	4	9	21	8	18	19	14	13	27	12	23	7	30	2	32	26	26	11	4	19	16	18	26	7	5	10	16	29
GREEN	19	12	13	15	14	15	19	14	13	11	7	17	17	19	15	16	16	15	13	10	15	12	16	13	15	15	13	17	18	13	12	21	16	10	16	18	16	22	18	10	12	10	14	13	17	10	19	10	12	11
LINKE	11	18	8	14	17	14	19	13	12	17	24	18	12	9	17	20	15	23	16	18	19	16	19	19	15	20	18	19	19	15	13	15	9	23	13	19	19	14	21	9	23	16	15	20	22	21	14	23	15	15
SPD	38	34	40	51	48	44	38	47	19	46	40	37	50	27	22	50	22	45	41	47	45	17	48	32	23	40	27	26	22	14	18	20	15	38	20	49	16	21	25	20	43	27	23	15	22	40	31	18	24	14
# Total Actors	126	187	125	189	270	668	200	472	103	323	106	135	569	101	176	137	212	192	306	503	135	179	139	167	458	195	121	115	123	162	144	148	127	158	957	209	272	167	160	161	266	130	100	146	146	143	106	309	125	479
# Actors in CDU	37	53	32	45	64	155	44	99	18	68	22	28	112	18	28	24	31	32	52	81	22	24	21	28	64	29	17	13	13	21	18	19	16	23	114	29	33	21	19	19	37	15	13	20	16	19	13	39	13	56

(c) CSU

Party \ First name	ALOIS	WALTRAUD	SIEGLINDE	SEPP	HEIDEMARIE	GUENTER	ANNELIESE	ALFONS	ELFRIEDE	JOSEF	HUBERT	MARGA	GERHARD	EDITH	KARLHEINZ	BRUNHILDE	WINFRIED	MARGIT	EDELTRAUD	REINHOLD	KORBINIAN	GÜNTER	WERNER	ELFI	HANNELORE	LUDWIG	ERWIN	ROSEMARIE	DIETER	WOLFGANG	RUDOLF	GÜNTHER	ALFRED	HERIBERT	FRANZ	HERBERT	IRMGARD	ANNEMARIE	ROSWITHA	LEONHARD	MANFRED	HARALD	GERLINDE	JÜRGEN	WALTER	BERNHARD	HELMUT	HILDEGARD	GISELA	HORST
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
AfD	39	25	28	43	26	30	24	38	26	38	32	26	37	26	38	36	31	26	22	34	18	40	38	31	26	31	41	25	39	38	38	40	41	35	37	38	26	22	28	20	41	40	25	42	39	34	39	25	26	40
CDU	10	10	11	10	8	9	12	14	7	12	9	12	10	10	8	9	10	7	12	12	11	12	12	11	11	11	9	10	12	12	11	9	12	17	11	11	11	12	7	6	10	10	8	11	10	11	11	11	10	12
CSU	51	49	47	47	46	45	44	43	43	43	43	43	43	43	42	42	42	42	42	42	41	41	41	41	41	40	40	40	40	40	40	40	40	39	39	39	39	39	39	39	39	38	38	38	38	38	38	38	38	38
FDP	10	8	8	7	9	15	9	14	2	13	15	7	14	12	11	6	16	10	9	14	24	13	15	10	11	15	10	11	14	18	15	13	10	11	14	13	8	12	8	20	15	16	9	16	14	15	14	11	9	14
GREEN	12	10	11	8	14	14	10	7	7	10	11	12	12	14	9	10	15	15	9	9	19	12	12	11	10	13	10	14	12	13	10	10	10	10	11	11	14	17	17	11	12	12	18	13	10	13	10	15	15	10
LINKE	10	10	12	9	20	14	13	10	12	12	13	8	15	14	14	13	14	15	15	17	9	17	17	14	17	15	11	17	18	19	14	16	14	9	15	15	10	13	15	15	18	17	14	18	15	14	18	12	15	18
SPD	11	25	18	8	22	18	24	17	20	14	17	24	18	20	17	19	24	17	21	16	18	20	19	20	22	18	21	23	21	21	16	18	20	20	15	18	27	19	21	11	20	20	20	20	20	16	20	22	24	19
# Total Actors	293	485	244	245	179	244	235	263	162	1803	707	113	2343	532	440	106	460	565	196	609	132	2059	3241	182	684	727	758	480	3754	6138	964	874	998	161	2379	1702	352	334	685	129	3769	2654	353	6435	1814	1755	2407	374	1241	2395
# Actors in CSU	104	174	85	87	57	76	76	80	60	552	218	37	672	162	135	33	128	180	63	177	39	546	859	54	202	206	214	139	956	1522	267	240	268	45	655	460	102	98	198	41	946	663	102	1558	474	472	609	107	342	600

(d) FDP

Party \ First name	JUSTUS	JULIUS	TILMAN	CONSTANTIN	NICOLAS	BENEDICT	CORNELIUS	MAGNUS	MAXIMILLIAN	PHILIPP	NIKLAS	KILIAN	HENRIK	MORITZ	HUBERTUS	MATS	NIKOLAS	JANIK	FREDERIK	BENEDIKT	THORBEN	FREDERIC	HENDRIK	MORTEN	HAUKE	DIETRICH	CARL	JULIAN	FELIX	HENNING	TILL	PHILIP	LORENZ	ARND	NICOLAI	THILO	LINUS	FABIAN	FYNN	LENNARD	JANNIK	LUKAS	NILS	MATHIS	CEDRIC	JANEK	SÖNKE	JONATHAN	TIMON	RUBEN
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
AfD	10	15	9	15	14	19	17	20	23	25	19	20	22	16	23	15	15	25	16	18	22	11	22	25	16	26	25	24	22	19	16	26	15	26	25	23	17	25	21	21	18	22	24	18	28	18	20	21	20	24
CDU	13	13	11	13	12	24	12	17	16	12	13	11	13	12	13	15	9	13	14	17	11	9	13	12	12	16	16	11	13	14	9	11	13	18	10	14	11	12	9	12	10	13	11	14	11	7	17	12	17	10
CSU	14	19	18	22	20	18	20	19	29	21	17	25	20	17	27	16	21	21	20	29	16	20	19	21	14	28	24	21	19	22	12	15	22	27	18	21	15	23	14	20	12	20	17	16	12	13	14	15	13	14
FDP	44	37	36	34	33	33	32	32	32	32	31	31	31	31	31	31	30	30	30	30	30	30	29	29	29	29	29	29	29	28	28	28	27	27	27	27	27	27	27	27	27	27	27	27	27	26	26	26	26	26
GREEN	14	16	23	15	14	15	17	16	14	15	14	14	14	19	11	16	14	8	17	16	12	16	12	13	17	19	11	15	16	17	17	14	21	12	13	16	24	15	14	17	14	16	15	23	11	21	22	19	18	12
LINKE	18	19	22	14	15	17	12	19	17	18	18	14	20	14	15	24	15	16	13	15	16	18	19	21	12	15	19	21	17	23	16	20	13	17	16	21	18	21	20	21	19	15	16	18	14	23	21	19	15	16
SPD	24	21	24	16	20	22	12	16	19	23	25	17	20	25	18	23	19	19	26	20	25	25	23	32	29	21	14	27	25	22	27	20	23	18	14	21	19	24	23	30	27	25	26	18	22	14	21	22	25	19
# Total Actors	304	843	177	394	696	176	170	272	2772	5028	2090	389	614	2005	168	123	185	137	542	957	234	266	1360	103	304	202	479	3276	4205	854	605	1077	387	144	218	411	266	3643	139	163	644	3957	2138	103	250	102	167	876	160	212
# Actors in FDP	97	225	45	105	180	39	45	63	598	1098	478	89	142	443	38	29	42	31	116	198	53	62	293	20	64	39	103	648	826	172	128	228	75	28	48	82	54	691	29	30	135	742	412	21	53	23	33	165	30	45

(e) GREEN

Party \ First name	RONJA	CLARA	HANNAH	FRIEDERIKE	AMELIE	NELE	JULE	MERLE	REBEKKA	PAULA	NORA	LINA	GRETA	LUCIA	KIRA	WIEBKE	LEONIE	SULTAN	HELEN	MIRI	JOHANNA	SOPHIE	TABEA	DEBORAH	LARA	SOPHIA	LEA	ANNIKA	LENA	KARO	MAREIKE	TIMKE	LOUISA	LAURA	HANNA	ESTHER	KLARA	LILLI	LEYLA	PAULINE	ELIF	CHARLOTTE	MAREN	MARLENE	PAULINA	REGINE	SUSANNA	MARA	ANNA	SVENJA
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
AfD	17	13	11	9	11	14	19	11	21	22	19	21	18	20	14	13	14	3	21	11	14	17	17	23	19	17	16	18	18	23	15	10	14	18	16	18	31	21	8	21	11	15	19	22	13	23	14	23	21	21
CDU	5	6	6	8	5	5	5	6	6	6	6	5	4	9	6	10	8	5	8	4	9	7	6	6	6	8	9	8	8	10	6	10	7	8	10	7	5	11	3	6	6	10	8	8	7	9	8	6	8	7
CSU	5	4	6	13	5	7	7	2	14	11	9	7	11	13	6	10	9	5	12	15	19	8	7	7	7	11	6	7	9	14	10	8	9	9	12	11	20	13	3	8	4	12	10	22	5	24	20	7	14	9
FDP	5	9	9	14	9	8	7	9	9	9	7	8	11	11	7	12	12	3	7	11	10	10	9	6	12	13	9	11	12	8	10	16	13	11	9	11	11	8	5	6	6	13	11	9	13	16	6	8	12	12
GREEN	43	40	40	40	38	36	35	35	34	33	33	33	33	33	33	33	33	33	32	32	32	32	32	32	32	32	31	31	31	31	31	30	30	30	30	30	29	29	29	29	29	29	29	29	29	28	28	28	28	28
LINKE	19	19	17	14	18	22	23	20	19	13	17	19	16	16	16	14	16	26	14	15	18	20	11	15	19	16	16	17	16	12	15	15	15	16	18	17	15	23	26	22	18	18	14	15	19	14	18	16	17	14
SPD	22	26	30	28	24	24	25	28	25	23	27	28	25	17	32	33	30	51	21	25	25	27	27	23	29	20	31	29	29	23	29	25	29	30	27	24	28	14	41	29	38	25	25	23	31	20	27	25	26	29
# Total Actors	220	299	656	295	163	261	440	137	164	418	377	374	118	170	198	308	420	100	195	109	1152	815	179	111	679	502	909	806	1582	132	376	121	177	2097	578	324	190	157	111	238	115	508	458	368	130	232	129	232	3702	613
# Actors in GREEN	83	102	220	93	56	81	130	43	43	119	106	103	33	47	57	81	113	26	55	31	287	213	52	31	175	136	240	206	399	34	99	32	46	509	140	81	40	39	28	58	30	120	113	83	32	49	30	58	830	144

(f) LINKE

Party \ First name	CEMAL	AZAD	YILMAZ	FIRAT	UMUT	JONA	GEORGIOS	ÖZGÜR	JANNES	DIMA	SERHAT	HÜSEYIN	CEM	CHRISTOS	KEMAL	FERHAT	BARIS	ERKAN	CIHAN	SERGEJ	EREN	MERLIN	LU	BÜLENT	DORIAN	CAN	MO	DENIZ	SINAN	HASAN	NI	JACOB	ALI	EL	NIKOLA	MAXIM	KAT	ERCAN	NIKOLAI	JAY	SAN	MOE	ARON	KRIS	FILIZ	SALIH	MEHMET	DI	ORHAN	ARI
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
AfD	8	8	9	15	17	14	24	9	16	40	15	5	13	12	9	12	15	18	12	39	15	23	31	12	21	17	23	15	13	10	34	17	12	37	18	30	27	13	19	33	34	21	29	29	8	15	11	36	13	24
CDU	3	7	6	4	6	9	8	2	11	3	4	5	5	6	4	2	4	3	7	3	5	10	7	4	11	8	6	6	8	8	9	9	12	8	9	10	8	1	8	8	9	5	7	12	7	8	5	9	6	6
CSU	4	6	9	6	4	13	17	5	10	5	5	6	5	21	6	8	7	5	6	14	9	11	19	8	17	8	16	9	7	8	22	12	8	18	14	7	16	8	23	15	16	20	10	22	8	3	7	24	6	19
FDP	5	6	5	6	8	17	7	9	19	7	9	5	10	10	2	6	4	7	8	9	10	19	14	5	26	13	16	9	4	7	15	15	6	8	9	19	10	8	19	13	12	20	13	17	5	7	8	12	7	9
GREEN	24	18	25	28	15	22	18	20	10	16	12	25	13	7	18	15	8	11	15	7	7	23	22	19	16	15	19	15	16	22	20	22	19	13	12	11	20	11	9	14	18	17	15	16	26	16	24	11	22	19
LINKE	44	43	42	40	38	37	37	37	35	35	35	34	34	34	34	33	33	33	33	32	32	32	31	31	31	31	30	30	30	30	30	30	30	30	29	29	29	29	29	29	28	28	28	28	28	28	28	28	28	28
SPD	39	41	45	44	37	27	26	40	30	9	40	47	45	27	43	50	44	42	41	12	48	18	17	44	27	43	23	41	46	45	32	25	43	21	23	19	22	50	19	17	21	23	22	22	47	40	47	30	46	26
# Total Actors	100	144	138	103	111	128	135	120	105	105	114	333	262	104	154	144	115	149	130	312	110	163	180	140	121	350	425	560	225	494	206	263	1740	553	114	169	187	120	187	248	106	123	102	189	110	101	755	207	139	104
# Actors in LINKE	35	48	41	29	34	34	36	36	28	32	33	88	71	30	45	38	33	41	35	86	28	38	40	35	25	80	98	136	55	115	38	60	402	119	29	39	41	29	43	55	22	26	23	36	24	24	161	38	30	22

(g) SPD

Party \ First name	TC	FATIH	RAMAZAN	MUHAMMED	GÖKHAN	MERT	METIN	SALIM	SERKAN	AYHAN	AHMET	HATICE	OSMAN	KENAN	SULTAN	YASIN	MUHAMMAD	BURAK	FERHAT	AYSE	FATMA	MOHAMMAD	SAID	UGUR	ERCAN	FARID	ISMAIL	HALIL	MURAT	ABDUL	MALIK	ENES	MOHAMAD	KADIR	CENGIZ	REZA	ÖMER	TOLGA	EREN	ABDULLAH	HÜSEYIN	HAKAN	YUSUF	AHMED	MUSTAFA	MOHAMED	KAAN	MEHMET	FILIZ	MESUT
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
AfD	8	16	13	14	14	19	6	17	14	14	11	4	17	9	3	22	10	19	12	6	7	9	7	11	13	6	8	11	10	8	15	22	11	15	14	11	12	17	15	10	5	15	10	14	11	14	21	11	8	8
CDU	4	4	5	8	7	3	6	9	4	2	7	4	7	10	5	7	31	5	2	2	6	24	21	6	1	14	9	6	5	17	15	7	30	5	2	19	7	6	5	14	5	7	7	27	11	20	3	5	7	3
CSU	5	10	11	8	8	4	6	10	5	14	10	6	7	6	5	10	7	7	8	8	3	7	5	8	8	10	8	9	9	9	8	5	8	9	11	8	11	6	9	7	6	11	7	10	9	8	5	7	8	7
FDP	5	11	4	6	8	7	8	6	7	8	6	6	5	6	3	7	3	7	6	4	5	4	2	9	8	6	4	7	5	2	10	6	0	11	9	4	7	6	10	5	5	9	8	4	6	10	8	5	6	5
GREEN	10	16	24	12	11	15	21	15	14	12	15	26	14	13	33	15	15	10	15	19	26	17	16	16	11	18	18	16	20	18	12	14	14	17	13	16	17	14	7	15	25	20	18	14	19	10	13	24	26	24
LINKE	15	14	17	10	22	18	27	16	22	17	25	18	21	24	26	17	14	19	33	25	19	12	20	25	29	25	25	18	26	19	16	11	17	21	23	19	18	23	32	22	34	25	26	13	25	18	23	28	28	26
SPD	63	58	57	56	56	55	55	54	54	53	52	52	52	51	51	51	51	50	50	50	50	50	50	49	49	49	49	49	48	48	48	48	48	48	48	48	48	48	47	47	47	47	47	47	47	47	47	47	47	46
# Total Actors	583	262	132	167	150	143	205	103	198	123	434	120	220	135	100	218	189	202	144	123	175	569	137	125	120	107	305	164	492	209	110	134	270	164	120	139	216	134	110	268	333	337	286	472	728	503	112	755	110	120
# Actors in SPD	336	118	58	82	67	65	87	44	89	55	180	54	93	58	41	86	73	86	57	54	75	230	56	50	50	41	123	69	191	84	43	57	102	63	48	54	86	53	42	105	123	119	109	172	269	194	43	271	40	46

mathematics (set theory), and analytics (social set analysis) but also to demonstrate its technical viability, suitability and utility by designing, developing and evaluating an IT-artifact, the Social Set Visualizer (SoSeVi). In other words, we postulated and - hopefully - illustrated that Set Theory in general is better suited from a mathematical standpoint to model human social associations than network theory or graph theory. Beyond the immediate social network and particularly on large scale social media platforms such as Facebook, Twitter and Tencent QQ, we believe, and hope, that this fundamental change in the foundational mathematical logic of the formal model from graphs to sets will allow for new insights.

C. Limitations

One of this paper's limitations is that we do not present domain-specific empirical findings in terms of political sciences and social media management. That said, such domain-specific empirical findings of the set theoretical approach can be found in [28], [29]. A second limitation is the lack of exposition of the full range of set theoretical approaches beyond the classical "crisp sets" discussed in the paper (for example: fuzzy sets, rough sets, random sets, Bayesian sets). A third and final limitation is the limited space devoted to the technical aspects of the IT-artifact. Also, the data set is only for 2017 and does not contain previous years of political discourse on Facebook.

D. Future Research

Current and planned future work in our Center for Business Data Analytics is addressing some of the theoretical limitations identified above in terms of developing formal models and analytical methods for fuzzy, rough and random sets. Furthermore, more advanced modeling of political social media discourse needs to be performed through machine learning. Our focus is on data visualization, and merging these capabilities with innovative methods of extracting meaningful insights from the social media data at hand. We suggest future work on the 2017 German federal election also takes into account not only the party Facebook pages, but also the Facebook pages of each individual member of parliament. This would enable analysis of further grass-roots political activity and discourse.

References

[1] W. S. Cleveland, "Data science: an action plan for expanding the technical areas of the field of statistics," *International Statistical Review*, vol. 69, no. 1, pp. 21–26, 2001. [Online]. Available: http://dx.doi.org/10.1111/j.1751-5823.2001.tb00477.x 1

[2] M. Loukides, *What Is Data Science?* O'Reilly Media, 2012. 1

[3] N. Ohsumi, "From data analysis to data science," in *Data Analysis, Classification, and Related Methods*. Springer Berlin Heidelberg, 2000, pp. 329–334. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-59789-3_52 1

[4] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabsi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne, "Computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009. 1

[5] R. Vatrapu, "Understanding social business." in *Emerging Dimensions of Technology Management*. Springer, 2013, pp. 147–158. 1

[6] J. Sterne, *Social media metrics: How to measure and optimize your marketing investment*. John Wiley & Sons, 2010. 1

[7] M. Sponder, *Social media analytics: effective tools for building, intrepreting, and using metrics*. McGraw-Hill, 2012. 1

[8] Z. Tufekci, "Big questions for social media big data: Representativeness, validity and other methodological pitfalls," *arXiv preprint arXiv:1403.7400*, 2014. 1

[9] J. L. Gross and J. Yellen, *Graph theory and its applications*. CRC press, 2005. 1

[10] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, "Network analysis in the social sciences," *Science*, vol. 323(5916), pp. 892–895, 2009. 1

[11] M. Emirbayer, "Manifesto for a relational sociology," *The American Journal of Sociology*, vol. 103(2), pp. 281–317, 1997. 1

[12] R. R. Mukkamala, A. Hussain, and R. Vatrapu, "Towards a formal model of social data," IT University of Copenhagen, Denmark, IT University Technical Report Series TR-2013-169, November 2013. 1

[13] ——, "Towards a set theoretical approach to big data analytics," in *3rd International Congress on Big Data (IEEE BigData 2014)*, June 2014. 1

[14] M. S. Mizruchi, "Social network analysis: Recent achievements and current controversies," *Acta sociologica*, vol. 37, no. 4, pp. 329–343, 1994. 1

[15] R. Vatrapu, R. R. Mukkamala, and A. Hussain, "A set theoretical approach to big social data analytics: Concepts, methods, tools, and findings," in *ECCS Satellite Workshop 2014*, 2014, pp. 22–24. 1, 2

[16] B. Flesch, R. Vatrapu, R. R. Mukkamala, and A. Hussain, "Social set visualizer: A set theoretical approach to big social data analytics of real-world events," in *Big Data (Big Data), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2418–2427. 1, 2

[17] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister, "Upset: visualization of intersecting sets," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1983–1992, 2014. 2, 3

[18] M. Berry, I. Garcia-Blanco, and K. Moore, "Press coverage of the refugee and migrant crisis in the eu: a content analysis of five european countries," 2016. 2

[19] S. A. Testa, "Financial (in) stability, banking crisis and policy implications: An empirical analysis on the eu countries," B.S. thesis, Università Ca'Foscari Venezia, 2017. 2

[20] J. Fritsch, "Frauen und führung in deutschland: Analyse der chancen und risiken der gesetzlichen frauenquote," 2016. 2

[21] A. Hussain and R. Vatrapu, "Social data analytics tool (sodato)," in *DESRIST 2014*, ser. Lecture Notes in Computer Science (LNCS). Springer, vol. 8463, 2014, pp. 368–372. 3

[22] ——, "Social data analytics tool: Design, development, and demonstrative case studies," in *Enterprise Distributed Object Computing Conference Workshops and Demonstrations (EDOCW), 2014 IEEE 18th International*, Sept 2014, pp. 414–417. 3

[23] A. Hussain, R. Vatrapu, D. Hardt, and Z. Jaffari, "Social data analytics tool: A demonstrative case study of methodology and software." in *Analysing Social Media Data and Web Networks*. Palgrave Macmillan, 2014. 3

[24] H. G. Miller and P. Mork, "From data to decisions: a value chain for big data," *IT Professional*, vol. 15, no. 1, pp. 57–59, 2013. 3

[25] Statista, "Mitgliederzahlen der politischen parteien in deutschland am 31. dezember 2016," Webpage, 2016, https://de.statista.com/statistik/daten/studie/1339/umfrage/mitgliederzahlen-der-politischen-parteien-deutschlands/. 5

[26] B. fr Politische Bildung, "Die soziale zusammensetzung der parteimitgliederschaften," Webpage, 2016, https://www.bpb.de/politik/grundfragen/parteien-in-deutschland/zahlen-und-fakten/140358/soziale-zusammensetzung. 6

[27] J. Mueller and G. Stumme, "Gender inference using statistical name characteristics in twitter," *arXiv preprint arXiv:1606.05467*, 2016. 6

[28] R. R. Mukkamala, J. I. Srensen, A. Hussain, and R. Vatrapu, "Detecting corporate social media crises on facebook using social set analysis," in *Proceedings of IEEE Bigdata Congress*, 2015. 10

[29] ——, "Social set analysis of corporate social media crises on facebook," in *Proceedings of IEEE 19th International Enterprise Distributed Object Computing Conference (EDOC)*, 2015. 10

Publication V

Social Interaction Model

Benjamin Flesch. *Social Interaction Model*. In Big Data (Big Data), 2018 IEEE International Conference on. IEEE, 2018

Social Interaction Model

Benjamin Johannes Flesch
Centre for Business Data Analytics
Copenhagen Business School, Denmark
bf.digi@cbs.dk

***Abstract*—This paper introduces a novel conceptual foundation for Social Set Analysis, the *Social Interaction Model*. It contributes to the state-of-the-art theory in Big Social Data by extending the existing *Social Data Model* through proposal of an improved concept which is formally grounded in set theory and relational algebra. The concept of *Interactions* based on one initial *Action* between two social media *Actors*, and zero or many *Reactions*, all referencing the initial *Action*, is presented. Furthermore, temporal and spatial dimensions are included as a core component of the proposed model, thus streamlining data analytics tasks in the realm of *Social Set Analysis*. Key differences between the existing *Social Data Model* and the *Social Interaction Model* contributed in this paper are the inclusion of non-textual artifact content types, the unification of a previously bipartite *Social Data Model*, the deprecation of an empirically vague notion of *Activities*, and improved interoperability between data sources.**

***Index Terms*—Big Data Analytics, Big Social Data, Computational Social Science, Social Set Analysis, Social Data Model.**

I. Introduction

In this paper I propose the *Social Interaction Model* (SIM), a generalized model of interactional social data, which extends upon the *Social Data Model* (SDM) [1], [4]. It simplifies the existing model through a set-based definition of interactions and the resulting artifacts. The model formalizes a two-dimensional framework based on location in space and time for Big Social Data Analytics. This provides additional empirical coherence with the application of Social Set Analysis (SSA) as presented in previous publications of our research group [5], [6]. SSA as discussed in this paper is concerned with the mobility of social actors across time and space. For comparison across dimension of time, we conduct SSA of big social data utilizing set theoretical intersections of various time periods along the time axis. Similarly, for comparison across location in space, we conduct set inclusions and exclusion based on the dimension of space. The SDM by Mukkamala et al. [1] is drawn from the theory of socio-technical interactions by Vatrapu [7]. A more detailed explication of the theoretical framework in terms of its ontological and epistemological assumptions and principles is beyond the scope of this paper but for details, please confer Vatrapu [7]. SDM was first utilized in the IT artifact of the Social Data Analytics Tool (SODATO) [8], [9]. Figure 1 illustrates the most recent version of the SDM, describing two concepts: *Interactions* (originally labeled *Social Graph* in first publication of SDM) and *Conversations* (originally labeled *Social Text*). Social Graph maps on to the first aspect of socio-technical interactions that involve perception and appropriation of affordances. Social Text maps on to the second aspect of socio-technical interactions. The model does not distinguish between an user and an actor. With respect to action/activity, an action is an atomic event done by an actor on an artifact, whereas an activity can spread across many actions, artifacts and actors. Social graph consists of the structure of the relationships emerging from the appropriation of social media affordances such as posting, linking, tagging, sharing and liking. It focuses on identifying the **actors** involved, the **actions** they take, the **activities** they undertake, and the **artifacts** they create and interact with. Social text consists of the communicative and linguistic aspects of the social media interaction such as the **topics** discussed, **keywords** mentioned, **pronouns** used and **sentiments** expressed.

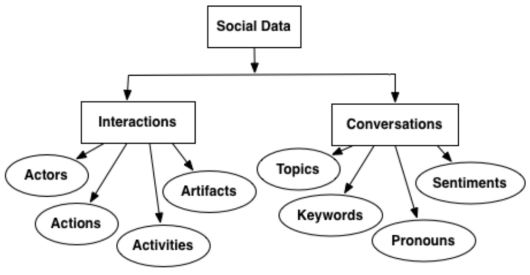


Fig. 1: Social Data Model (est. 2013, updated 2016) [1]–[3]

II. Conceptual Model

Figure 2 presents the conceptual foundation of the *Social Interaction Model* (SIM). The conceptual model emerged during research on the Social Set Visualizer [10]–[12], a visual analytics software tool implementing SSA. The proposed model distinguishes between three major components of social data: ***Actors***, ***Interactions*** and ***Artifacts***. ***Actors*** depict any kind of user or entity that can be interacted with in the realm of social media. Each ***Actor*** consists of a unique ***Location in Space*** and a set of ***Artifacts*** which depict the actor's attributes, e.g. a profile picture or bio text. An ***Interaction*** depicts a set of one initial ***Action*** and zero or many ***Reactions*** that respond to the initial ***Action*** or one of its ***Reactions***. ***Actions*** always occur between two ***Actors***, one originating ***Actor*** and one receiving ***Actor***. ***Reactions*** always originate from one ***Actor*** and are targeted at another ***Action*** or ***Reaction***. The context of each ***Interaction*** consists of the ***Location*** in ***Space*** and ***Time***, thus allowing analytics along space and time dimensions. ***Location in Space*** is provided by ***Actors*** who are part of the initial ***Action***, as by definition every ***Actor*** has a

certain location in space. Therefore, the ***Location in Space*** for the whole interaction is set by the location attributes of the initial ***Action*** of each ***Interaction***. ***Location in Time*** is attached to every single ***Action*** and ***Reaction*** as a conventional time stamp. Every individual ***Action*** or ***Reaction*** results in creation of a set of ***Artifacts***. These ***Artifacts*** can be aggregated on the ***Interaction*** level, resulting in a set of all ***Artifacts*** created during a certain ***Interaction***. ***Artifacts*** depict any user-generated content, such as text posts, emotions, or media uploads. ***Artifacts*** consists of a certain content type and a user-generated payload. For theoretical fidelity with existing SDM, the SIM also distinguishes between three content types, *Social Videos*, *Social Images* and *Social Text*. *Artifacts* can be analyzed to extract further information om *Topics*, *Keywords*, *Sentiments*, and *Pronouns*.

III. Formal Model

Definition 1. We define l as a certain location in space expressed by a three-dimensional vector containing the *data source*, *data type* and an *unique identifier* (ID). l depicts *Location in space*. We define t as a certain point in time expressed through a timestamp. t depicts *Location in time*.

Definition 2. We define a as a tuple $a = (c, ct)$, where c is data of a certain content type $ct \in \mathbb{R}$. $\mathbb{R}$ is the set of all content types.

> a depicts an $Artifact$. The $Artifact$ contains data c of a certain content type ct with $ct \in \mathbb{R}$.

Definition 3. We define α as a tuple $\alpha = (l,\ A_\alpha)$ with l as a location in space, and A_α as a set of artifacts a who are direct attributes of α. We define Θ as the set of all $Actors$ α.

> α depicts an $Actor$. It describes a point in space l with attributes of the *Actor* encoded as a set of $Artifacts$ A_α.

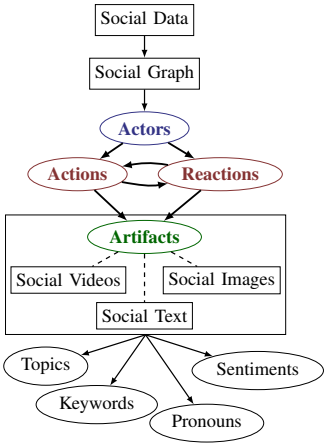


Fig. 2: Concept of proposed Social Interaction Model.

Definition 4. We define δ as the $Type$ of a certain $Action$, and Δ as the set of all $Action$ $Types$ δ. We define $\beta = \beta_{\delta,\alpha_1,\alpha_2}$ as a relation $\beta_{\delta,\alpha_1,\alpha_2} : (\delta, \alpha_1, \alpha_2) \rightarrow (t,\ A_\beta)$ of $Type$ $\delta \in \Delta$ originating from $Actor$ $\alpha_1 \in \Theta$ to $Actor$ $\alpha_2 \in \Theta$, resulting in a tuple $(t,\ A_\beta)$ where t is a certain point in time and A_β is a set of artifacts a created by the relation β.

> $\beta_{\delta,\alpha_1,\alpha_2}$ depicts an $Action$ of $Type$ δ originating from $Actor$ α_1 directed at α_2.

Definition 5. We define $\beta_{1_{\delta,\alpha,\beta_2}}$ as a relation $\beta_{1_{\delta,\alpha,\beta_2}} : (\delta, \alpha, \beta_2) \rightarrow (t, A_{\beta_1})$ of $Type$ $\delta \in \Delta$ originating from actor $\alpha \in \Theta$ to relation β_2, resulting in a tuple (t, A_{β_1}) with t and A_{β_1} analogous to definition 4.

> $\beta_{\delta,\alpha,\beta_2}$ depicts a $Reaction$ of $Type$ δ originating from $Actor$ α directed at previous $Action$ β_2, with β_2 depicting either an $Action$ $\beta_{\delta,\alpha_1,\alpha_2}$ (def. 4) or a $Reaction$ $\beta_{\delta,\alpha,\beta_2}$ (def. 5). Both $Actions$ and $Reactions$ β return a tuple (t, A_β) containing timestamp t and a set of $Artifacts$ A_β.

Definition 6. We define $I_{\beta_{\delta,\alpha_1,\alpha_2}}$ as the set of all relations β containing a relation $\beta_{\delta,\alpha_1,\alpha_2}$ of $Type$ $\delta \in \Delta$ between two $Actors$ $\alpha_1, \alpha_2 \in \Theta$, and all other relations β either referencing that certain relation or any other relation β in $I_{\beta_{\delta,\alpha_1,\alpha_2}}$. We recursively define $I_\beta = I_{\beta_{\delta,\alpha_1,\alpha_2}} = \{\beta_{\delta,\alpha_1,\alpha_2}\} \cup \{\beta_{\delta_1,\alpha,\beta_2} | \exists \delta_1 \in \Delta, \alpha \in \Theta : \beta_2 \in I_{\beta_{\delta,\alpha_1,\alpha_2}}\}$.

> I_β depicts an $Interaction$. It consists of a certain $Action$ $\beta = \beta_{\delta,\alpha_1,\alpha_2}$ between two $Actors$ α_1 and α_2 and all $Reactions$ $\beta_{\delta,\alpha,\beta_2}$ that reference it. $Reactions$ to $Reactions$ are also included.

Definition 7. We define $S_\alpha = \{I_{\beta_{\delta,x,\alpha}} | \forall x \in \Theta, \delta \in \Delta\}$ as the set of all $Interactions$ with $Actor$ α. We define $S = \{S_\alpha | \forall \alpha \in \Theta\}$ as our corpus of interactional data. Figure 3 illustrates formal definition of the *Social Interaction Model*.

IV. Discussion

We compare the SIM with the existing SDM and present key differences. **Interactions:** In applied SSA, a notion of interactions is needed, consisting of a data structure comparable to a linked, timestamped list of one initial *Action* performed by a social media *Actor* directed at another *Actor* and zero or many *Reactions* to that initial action. SIM incorporates Interactions as core principle and goes further than the SDM in that regard. **Definition of Temporal and Spatial Dimensions:** To formalize SSA methodology, two dimensions, one for time and one for space, are required. The SIM extends the existing SDM to support these two dimensions of *Location in Space* and *Location in Time*. **Non-Textual Artifacts:** The existing SDM lacks support for non-textual *Artifact* content

types apart from *Social Text*, such as images and videos within social data. This is rectified by the proposed SIM through addition of *Social Images* and *Social Videos*, from which meaningful information can be extracted by utilizing state-of-the-art computational approaches. Therefore, *Topics*, *Keywords*, *Pronouns* and *Sentiments* should not be attached to *Social Text* domain as theorized in the SDM, but attached to *Artifacts* of any content type as proposed by the SIM (see figure 2). **Unification of Bipartite SDM:** The proposed SIM unifies the bipartite SDM into one coherent concept of a set-based definition of the social data and the social graph. This step is straightforward after a slight refinement in the definition of *Artifacts*. This modification enables the SIM to express that all meaning such as *Topics*, *Keywords*, *Pronouns*, and *Sentiments* can be extracted from the *Artifact* data. With *Social Text* being a specific type of *Artifact* data, we observe that there is no inherent conflict between both models, but that SIM is a logical extension and generalization of the existing SDM. **Deprecation of "Activities":** *Activities* as defined in the existing SDM are a vague concept without clear mapping to real-world Big Social Data used in SSA. The formal definition in the original publication concerns a mapping function from *Artifacts* to *Activities*, with the presented example being that *Activity* is a "promotion" of products by a clothing retailer on Facebook [1], that may span many Actors, Actions and Artifacts. Thus, the notion of *Activity* in the SDM apparently tries to capture the goal or intention of a social media *Actor* that is behind their *Action* to broadcast a certain *Artifact* to the social network. I argue that this non-public information on goal or intention of the *Actor* is empirically difficult to obtain both for researchers and the social network operators. Therefore, due to empirical difficulties in acquiring this data with sufficient precision, and the proven precedent of only being able to guess the intention of an *Activity*, the concept of *Activities* has not been included in the SIM. **Improved Interoperability between Data Sources:** Based on the dimension of *Location in Space* which is included in the proposed SIM, it is conceptually possible to interoperate between *Artifacts* from different social media data sources in line with the model

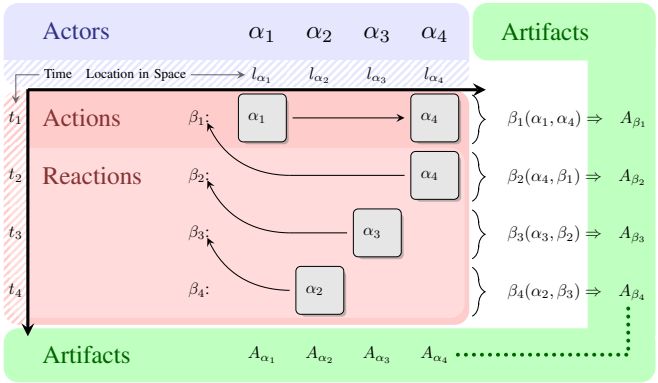


Fig. 3: Illustration of proposed *Social Interaction Model* based on Actors, Interactions, and Artifacts.

definition. For example, after a dimensionality reduction to the temporal dimension *Location in Time*, *Artifacts* from multiple data sources such as Twitter and Facebook can be grouped and compared for SSA purposes.

V. Conclusion

In this paper, I have contributed a new conceptual foundation for Social Set Analysis, the *Social Interaction Model*, based on a set-theoretical formal model which introduces the notion of *Interactions* based on one initial *Action* and zero or many *Reactions*. I have articulated further key differences to the existing *Social Data Model*, such as the definition of temporal and spatial dimensions, the inclusion of non-textual *Artifact* content types, the unification of a previously bipartite model, improved interoperability between big social data sources, and the deprecation of the empirically vague *Activity* concept. On top of that, the proposed *Social Interaction Model* simplifies the set-based formalization of the *Social Data Model*, therefore improving the state-of-the-art in Big Social Data by providing a streamlined conceptual model for utilization in real world use cases. Temporal and spatial dimensions are introduced as core components of the proposed model, thereby a guideline and standardization for Social Set Analysis tasks is contributed to the research community.

References

[1] R. R. Mukkamala, A. Hussain, and R. Vatrapu, "Towards a formal model of social data," *IT University of Copenhagen, Denmark, IT University Technical Report Series TR-2013-169*, vol. 1, no. 3, p. 5, 2013.

[2] R. Mukkamala, A. Hussain, and R. Vatrapu, "Fuzzy-set based sentiment analysis of big social data," in *Enterprise Distributed Object Computing Conference (EDOC), 2014 IEEE 18th International*, pp. 71–80, 2014.

[3] R. Vatrapu, R. R. Mukkamala, A. Hussain, and B. Flesch, "Social set analysis: A set theoretical approach to big data analytics.," *Ieee Access*, vol. 4, pp. 2542–2571, 2016.

[4] R. R. Mukkamala, A. Hussain, and R. Vatrapu, "Fuzzy-set based sentiment analysis of big social data," in *18th Intl. Enterprise Distributed Object Computing Conference (EDOC 2014)*, pp. 71–80, IEEE, 2014.

[5] R. Vatrapu, R. R. Mukkamala, and A. Hussain, "A set theoretical approach to big social data analytics: Concepts, methods, tools, and findings," in *ECCS Satellite Workshop 2014*, pp. 22–24, 2014.

[6] R. Vatrapu, A. Hussain, N. B. Lassen, R. R. Mukkamala, B. Flesch, and R. Madsen, "Social set analysis: four demonstrative case studies," in *Proceedings of the 2015 International Conference on Social Media & Society*, p. 3, ACM, 2015.

[7] R. K. Vatrapu, "Explaining culture: An outline of a theory of socio-technical interactions," in *Proceedings of the 3rd International Conference on Intercultural Collaboration*, ICIC '10, (New York, NY, USA), pp. 111–120, ACM, 2010.

[8] A. Hussain and R. Vatrapu, "Social data analytics tool (sodato)," in *DESRIST 2014*, vol. 8463 of *Lecture Notes in Computer Science (LNCS). Springer*, pp. 368–372, 2014.

[9] A. Hussain and R. Vatrapu, "Social data analytics tool: Design, development, and demonstrative case studies," in *Enterprise Distributed Object Computing Conference Workshops and Demonstrations (EDOCW), 2014 IEEE 18th International*, pp. 414–417, 2014.

[10] B. Flesch, R. Vatrapu, R. R. Mukkamala, and A. Hussain, "Social set visualizer: A set theoretical approach to big social data analytics of real-world events," in *Big Data (Big Data), 2015 IEEE International Conference on*, pp. 2418–2427, IEEE, 2015.

[11] B. Flesch, R. R. Mukkamala, A. Hussain, and R. Vatrapu, "Social set visualizer (sosevi) ii: Interactive social set analysis of big data.," in *SetVR@ Diagrams*, pp. 19–28, 2016.

[12] B. Flesch, A. Hussain, and R. Vatrapu, "Social set visualizer: Demonstration of methodology and software," in *2015 IEEE 19th International Enterprise Distributed Object Computing Workshop*, pp. 148–151, 2015.

Appendix A

Literature Review Visual Analytics

Article	Data Type	Data Collection	Data Set Attributes	Application Domain	Data Types Visualized	Visualization Techniques	User Evaluation
[Zimmerman *et al.* 2014]	Twitter, Facebook	Radian6, SODATO	20000 Facebook posts	Marketing	Emotion scores from classifier	Scatter plot	None
[?]	Twitter, Facebook, Youtube, Instagram, Tumblr	Radian6, SODATO	Dynamic	Marketing	Topic and story performance	Scatter plot / radial plot	5 clients, no task specified
[Giatsoglou *et al.* 2016]	Twitter, Flickr, Foursquare	Custom	65348 geo-tagged tweets, 7500 photos	Smart City	Content clustering, trending topic detection and event detection	Map with overlay, pie chart	Case study with city of Santander, Spain
[Heer & Agrawala 2007]	NA	NA	NA	Collaborative analytics	Various	Various	Best practice evaluation
[Yeon *et al.* 2016]	Twitter, Newspapers	Custom	53M tweets and 150k news documents	Abnormal events / news	Content clustering, trend prediction	Map, calendar timeline, bar charts	8 students, using tool to find unusual events
[Chae *et al.* 2014]	Twitter	Custom	2.2M geotagged tweets within the US (per day)	Abnormal events / news	Spatiotemporal analysis	Spatiotemporal visualization w/ maps and overlays	None
[Kim *et al.* 2016a]	Twitter	Custom	230k tweets, of which 24k are usable for sentiment analysis	Marketing	Disparity in volume, intention and sentiment	Line charts	None

Article	Data Type	Data Collection	Data Set Attributes	Application Domain	Data Types Visualized	Visualization Techniques	User Evaluation
[Marcus *et al.* 2011]	Twitter	Custom	NA	Abnormal events / news	Sentiment, volume and peak detection	Sentiment timeline, maps	12 users + structured interviews with investigative journalists
[Padmanabhan *et al.* 2014]	Twitter	CyberGIS software environment	15M tweets on a 7-day sliding window	Disease tracking	Interactive spatio-temporal exploration, trajectory mapping	Maps with heatmap overlays	None
[Yang *et al.* 2016]	Twitter	Custom, GIS software	>2M tweets	Tweet tracking	Tweets vs. population size in selected areas	Dashboard, heatmaps, word clouds	User interface evaluation w/o any test subjects
[Ribarsky *et al.* 2014]	Twitter	Custom	20B tweets	Customer analytics	Event bursts, topic modeling	Timeline, event shapes, line graphs	Interviews with business partners
[Chae 2015]	Twitter	Custom	22k tweets	Supply chain research	General activity, sentiment	SNA graph, timeline, sentiment line chart	None
[Carley *et al.* 2014]	Twitter, LexisNexis news data	TweetTracker, LexisNexis	30M tweets, 10k news articles	Political science	Ceneral twitter activity, topic networks	Line charts, topic graphs	None

Article	Data Type	Data Collection	Data Set Attributes	Application Domain	Data Types Visualized	Visualization Techniques	User Evaluation
[Cheng & Edwards 2015]	Sina Weibo posts	Manual analysis of repost behavior of most popular posts by CCTV account	12k posts	Tourism	Geo-visual analysis of reposts, repost volume, gender distribution	Word clouds, volume charts, bar charts, distribution plot, maps	Self-study with 4 posts
[Kucher *et al.* 2015]	Twitter, Facebook	Cavagai.se for social text and sentiment	15M Tweets per day (min. 10 days), 5k Facebook posts	Linguistics	Stance and sentiment model	Dashboard with timeline chart; document analysis interface; bubble plot	Expert review with 2 experts, feedback on analysis workflow and tool design
[Ma *et al.* 2016]	Large-scale video content	Movies	Movie "The Matrix" with 14 main characters and 76 characters total, 14 main events and 83 kinds of connections	Explorative Analysis of Video Content	video content; associations between video content; temporal / spacial / associations of video content	Map; user annotations / gesture-based commands; POI labeled map; video storylines	18 participants for evaluation of interface and video exploration time
[Quinn *et al.* 2016]	Twitter	DataSift	350k tweets	Protest analysis	Sentiment, volume, hashtag popularity, network analysis	Bar charts, distribution, non-interactive	None

Article	Data Type	Data Collection	Data Set Attributes	Application Domain	Data Types Visualized	Visualization Techniques	User Evaluation
[Ramanathan *et al.* 2013]	Twitter, Health insurance claims data	NA	NA	Public health surveillance	Spatial and temporal patterns during the 2009-2010 pandemic H1N1 flu season using claims data	Map with overlays, pie charts, line charts, heatmaps	None
[Chua *et al.* 2015]	Twitter (geotagged posts)	NA	750k tweets	Mobility study	Spatial planning and urban flow analysis, calculation of pathways of urban flows	Map with flow vector overlays, timeline, histograms	None
[Magdy *et al.* 2014]	Twitter (geotagged posts)	Taghreed indexer with twitter firehose	>5k tweets per second from gulf region	Twitter monitoring	Tweets, metadata, geospatial data, language analysis	Interactive dashboard, maps, line graphs, pie charts, word clouds (non-alphabetic; colorful), timelines	None
[Pääkkö-nen 2016]	Twitter, Facebook	Twitter Firehose	NA	Database systems	Database performance metrics	Bar charts, line charts	None, only software benchmarks

Article	Data Type	Data Collection	Data Set Attributes	Application Domain	Data Types Visualized	Visualization Techniques	User Evaluation
[Pfeffer *et al.* 2015]	Twitter	DOLLY project	300k tweets (geocoded data since 2012 in predefined bounding boxes)	Development research	Tweet density, distribution across resorts, frequency of security-related issues	Tables, heatmap overlay on city maps	None
[Xu *et al.* 2016]	Sina Weibo, websites	NA	1.5M websites, 150 events	Urban emergency event analysis	Event evolution, event burst metrics, event membership	Line charts	None
[Wei *et al.* 2016]	Twitter	Custom	73M tweets	Social network analysis	Temporal change in number of edges and nodes in twitter graph per country; language distribution; network diameter; various SNA metrics	Bar chart; heatmaps; maps; percentage bar chart; scatter plots	None

Article	Data Type	Data Collection	Data Set Attributes	Application Domain	Data Types Visualized	Visualization Techniques	User Evaluation
[ben Khalifa *et al.* 2016]	Twitter	Custom	500k tweets	Crowd detection	Number of calculated clusters of twitter users, tweets per cluster, other cluster metrics	Map with cluster overlay, tables, line charts, scatter plot, histogram	None
[Archambault & Hurley 2014]	Telco subscriber data and call detail records	Proprietary telco data	Telco analytics graph with 1Bn edges	Social network analysis	Churn rate, customer metrics	Pixel-oriented display technique with interactive controls	With project partners only
[Benjamin *et al.* 2014]	Extermist web forums	Dark Web Forum Portal (DWFP) project which collects posts from extremism forums	500k messages from 15k users	Text authorship	Authorship styles, sentiment, use of bad words	Table, heatmaps, radar charts	31 participants to evaluate performance of visualization tool
[Liu *et al.* 2014]	NA	NA	NA	Data visualization	Empirical methodologies, interactions, frameworks, and applications; storylines, genes, graphs	Graphs, matrix, tables, dashboards, flow charts, histograms, d3.js-based visualizations, compressed adjacency matrices	None

Article	Data Type	Data Collection	Data Set Attributes	Application Domain	Data Types Visualized	Visualization Techniques	User Evaluation
[Vorvoreanu *et al.* 2013]	Facebook, Twitter	Radian6	71k tweets and 14k Facebook posts with sentiment	Marketing analytics	Sentiment results, post volume, event timeline	Pie chart, word cloud, line chart	None
[Nam *et al.* 2015]	Blogs, Twitter, Facebook	Custom	23k Tweets and 23k Facebook posts	Social network analysis	Linear model results, degree centralities, web ecology	Table, network diagrams	None
[Ferrara 2012]	Facebook	Custom	16M Facebook users	Social network analysis	Community size and metrics, community detection results	Scatter plot, hierarchy-based circular visualization algorithm	None
[Jha *et al.* 2016]	Facebook	Custom, manual	2.6k Facebook posts across 26 state health departments	Public health surveillance	Number of posts per category, per department	Scatter plots	None
[Muelder *et al.* 2014]	Facebook, Twitter	NA	3M posts from MedHelp	Social Science and Sociology	Entities of groups, sets of relationships between entities	Network diagrams, graphs, treemaps, timelines, storylines	None

Article	Data Type	Data Collection	Data Set Attributes	Application Domain	Data Types Visualized	Visualization Techniques	User Evaluation
[Li *et al.* 2016]	DBLP, Facebook, Youtube, Orkut	SNAP (Leskovec and Krevl 2014) data library for analyzing large information networks	4k users from Facebook, 1.1M from YT, 3M from Orkut	Social network analysis	Social network graph abstractions based on community aggregation	Network graphs, 3d scatter plot	None
[Cvijikj & Michahelles 2013]	Facebook	Custom	205k Facebook posts	Marketing Analytics / Social Media Marketing	Post distribution, model results for engagement measures, brand likes and page growth	Bar charts, tables	None
[Kim *et al.* 2016b]	Twitter	Custom	All tweets for Reuters top news account	E-Learning / Social Learning	Results of news classification	Matrix visualization, tables, bar charts	30 students evaluate classification accuracy, user clustering, satisfaction

Article	Data Type	Data Collection	Data Set Attributes	Application Domain	Data Types Visualized	Visualization Techniques	User Evaluation
[Acharya & Park 2016]	Blogs and NGO websites	Webometics Analysis, Bing API	50 seed websites	Development research / Social network analysis	NGO links and metions, link impact analysis, descriptive statiustcs, network analysis	Tables, line charts, network graph	None
[Lee *et al.* 2016]	Twitter, Blogs	Custom	600k tweets	Cyber security	Descriptive statistics, topic detection, topic distribution, clustering results	Network graph, tables, bar chart, scatter plot	None
[See-To & Ngai 2016]	TMall.com transaction and review data	Export from TMall	10k transactions and 9,5k customer reviews in 2,5k products in 4 shops all in certain 30-day period	Supply chain forecasting	product demand fit, OLS regression model results	Scatter plot	None

Article	Data Type	Data Collection	Data Set Attributes	Application Domain	Data Types Visualized	Visualization Techniques	User Evaluation
[Dos Santos Jr *et al.* 2016]	Twitter and GDELT data related to violent events	Custom, GDELT	, 10M tweets, 10M GDELT events in 2011 middle-east war	Event visualization	Spatio-temporal storytelling, structured violent event timelines, storyline set aggregations, spatio-logical inference model	Map with overlay, timlines, stylized cause and effect relationships, graphs, line charts	none
[Sun *et al.* 2013]	Twitter, Research co-authorship	Various	NA	Data visualization	Topic evolution, topic competition	Scatter plots, event timelines, graphs	None

TITLER I PH.D.SERIEN:

2004

1. Martin Grieger
Internet-based Electronic Marketplaces and Supply Chain Management

2. Thomas Basbøll
LIKENESS
A Philosophical Investigation

3. Morten Knudsen
Beslutningens vaklen
En systemteoretisk analyse of moderniseringen af et amtskommunalt sundhedsvæsen 1980-2000

4. Lars Bo Jeppesen
Organizing Consumer Innovation
A product development strategy that is based on online communities and allows some firms to benefit from a distributed process of innovation by consumers

5. Barbara Dragsted
SEGMENTATION IN TRANSLATION AND TRANSLATION MEMORY SYSTEMS
An empirical investigation of cognitive segmentation and effects of integrating a TM system into the translation process

6. Jeanet Hardis
Sociale partnerskaber
Et socialkonstruktivistisk casestudie af partnerskabsaktørers virkelighedsopfattelse mellem identitet og legitimitet

7. Henriette Hallberg Thygesen
System Dynamics in Action

8. Carsten Mejer Plath
Strategisk Økonomistyring

9. Annemette Kjærgaard
Knowledge Management as Internal Corporate Venturing
– a Field Study of the Rise and Fall of a Bottom-Up Process

10. Knut Arne Hovdal
De profesjonelle i endring
Norsk ph.d., ej til salg gennem Samfundslitteratur

11. Søren Jeppesen
Environmental Practices and Greening Strategies in Small Manufacturing Enterprises in South Africa
– A Critical Realist Approach

12. Lars Frode Frederiksen
Industriel forskningsledelse
– på sporet af mønstre og samarbejde i danske forskningsintensive virksomheder

13. Martin Jes Iversen
The Governance of GN Great Nordic
– in an age of strategic and structural transitions 1939-1988

14. Lars Pynt Andersen
The Rhetorical Strategies of Danish TV Advertising
A study of the first fifteen years with special emphasis on genre and irony

15. Jakob Rasmussen
Business Perspectives on E-learning

16. Sof Thrane
The Social and Economic Dynamics of Networks
– a Weberian Analysis of Three Formalised Horizontal Networks

17. Lene Nielsen
Engaging Personas and Narrative Scenarios – a study on how a user-centered approach influenced the perception of the design process in the e-business group at AstraZeneca

18. S.J Valstad
Organisationsidentitet
Norsk ph.d., ej til salg gennem Samfundslitteratur

19. Thomas Lyse Hansen
Six Essays on Pricing and Weather risk in Energy Markets

20. Sabine Madsen
Emerging Methods – An Interpretive Study of ISD Methods in Practice

21. Evis Sinani
The Impact of Foreign Direct Investment on Efficiency, Productivity Growth and Trade: An Empirical Investigation

22. Bent Meier Sørensen
Making Events Work Or, How to Multiply Your Crisis

23. Pernille Schnoor
Brand Ethos
Om troværdige brand- og virksomhedsidentiteter i et retorisk og diskursteoretisk perspektiv

24. Sidsel Fabech
Von welchem Österreich ist hier die Rede?
Diskursive forhandlinger og magtkampe mellem rivaliserende nationale identitetskonstruktioner i østrigske pressediskurser

25. Klavs Odgaard Christensen
Sprogpolitik og identitetsdannelse i flersprogede forbundsstater
Et komparativt studie af Schweiz og Canada

26. Dana B. Minbaeva
Human Resource Practices and Knowledge Transfer in Multinational Corporations

27. Holger Højlund
Markedets politiske fornuft
Et studie af velfærdens organisering i perioden 1990-2003

28. Christine Mølgaard Frandsen
A.s erfaring
Om mellemværendets praktik i en transformation af mennesket og subjektiviteten

29. Sine Nørholm Just
The Constitution of Meaning – A Meaningful Constitution?
Legitimacy, identity, and public opinion in the debate on the future of Europe

2005

1. Claus J. Varnes
Managing product innovation through rules – The role of formal and structured methods in product development

2. Helle Hedegaard Hein
Mellem konflikt og konsensus – Dialogudvikling på hospitalsklinikker

3. Axel Rosenø
Customer Value Driven Product Innovation – A Study of Market Learning in New Product Development

4. Søren Buhl Pedersen
Making space
An outline of place branding

5. Camilla Funck Ellehave
Differences that Matter
An analysis of practices of gender and organizing in contemporary workplaces

6. Rigmor Madeleine Lond
Styring af kommunale forvaltninger

7. Mette Aagaard Andreassen
Supply Chain versus Supply Chain Benchmarking as a Means to Managing Supply Chains

8. Caroline Aggestam-Pontoppidan
From an idea to a standard
The UN and the global governance of accountants' competence

9. Norsk ph.d.

10. Vivienne Heng Ker-ni
An Experimental Field Study on the

Effectiveness of Grocer Media Advertising
Measuring Ad Recall and Recognition, Purchase Intentions and Short-Term Sales

11. Allan Mortensen
Essays on the Pricing of Corporate Bonds and Credit Derivatives

12. Remo Stefano Chiari
Figure che fanno conoscere
Itinerario sull'idea del valore cognitivo e espressivo della metafora e di altri tropi da Aristotele e da Vico fino al cognitivismo contemporaneo

13. Anders McIlquham-Schmidt
Strategic Planning and Corporate Performance
An integrative research review and a meta-analysis of the strategic planning and corporate performance literature from 1956 to 2003

14. Jens Geersbro
The TDF – PMI Case
Making Sense of the Dynamics of Business Relationships and Networks

15 Mette Andersen
Corporate Social Responsibility in Global Supply Chains
Understanding the uniqueness of firm behaviour

16. Eva Boxenbaum
Institutional Genesis: Micro – Dynamic Foundations of Institutional Change

17. Peter Lund-Thomsen
Capacity Development, Environmental Justice NGOs, and Governance: The Case of South Africa

18. Signe Jarlov
Konstruktioner af offentlig ledelse

19. Lars Stæhr Jensen
Vocabulary Knowledge and Listening Comprehension in English as a Foreign Language
An empirical study employing data elicited from Danish EFL learners

20. Christian Nielsen
Essays on Business Reporting
Production and consumption of strategic information in the market for information

21. Marianne Thejls Fischer
Egos and Ethics of Management Consultants

22. Annie Bekke Kjær
Performance management i Proces-innovation
– belyst i et social-konstruktivistisk perspektiv

23. Suzanne Dee Pedersen
GENTAGELSENS METAMORFOSE
Om organisering af den kreative gøren i den kunstneriske arbejdspraksis

24. Benedikte Dorte Rosenbrink
Revenue Management
Økonomiske, konkurrencemæssige & organisatoriske konsekvenser

25. Thomas Riise Johansen
Written Accounts and Verbal Accounts
The Danish Case of Accounting and Accountability to Employees

26. Ann Fogelgren-Pedersen
The Mobile Internet: Pioneering Users' Adoption Decisions

27. Birgitte Rasmussen
Ledelse i fællesskab – de tillidsvalgtes fornyende rolle

28. Gitte Thit Nielsen
Remerger
– skabende ledelseskræfter i fusion og opkøb

29. Carmine Gioia
A MICROECONOMETRIC ANALYSIS OF MERGERS AND ACQUISITIONS

30. Ole Hinz
Den effektive forandringsleder: pilot, pædagog eller politiker?
Et studie i arbejdslederes meningstilskrivninger i forbindelse med vellykket gennemførelse af ledelsesinitierede forandringsprojekter

31. Kjell-Åge Gotvassli
Et praksisbasert perspektiv på dynamiske
læringsnettverk i toppidretten
Norsk ph.d., ej til salg gennem Samfundslitteratur

32. Henriette Langstrup Nielsen
Linking Healthcare
An inquiry into the changing performances of web-based technology for asthma monitoring

33. Karin Tweddell Levinsen
Virtuel Uddannelsespraksis
Master i IKT og Læring – et casestudie i hvordan proaktiv proceshåndtering kan forbedre praksis i virtuelle læringsmiljøer

34. Anika Liversage
Finding a Path
Labour Market Life Stories of Immigrant Professionals

35. Kasper Elmquist Jørgensen
Studier i samspillet mellem stat og erhvervsliv i Danmark under
1. verdenskrig

36. Finn Janning
A DIFFERENT STORY
Seduction, Conquest and Discovery

37. Patricia Ann Plackett
Strategic Management of the Radical Innovation Process
Leveraging Social Capital for Market Uncertainty Management

2006

1. Christian Vintergaard
Early Phases of Corporate Venturing

2. Niels Rom-Poulsen
Essays in Computational Finance

3. Tina Brandt Husman
Organisational Capabilities, Competitive Advantage & Project-Based Organisations
The Case of Advertising and Creative Good Production

4. Mette Rosenkrands Johansen
Practice at the top
– how top managers mobilise and use non-financial performance measures

5. Eva Parum
Corporate governance som strategisk kommunikations- og ledelsesværktøj

6. Susan Aagaard Petersen
Culture's Influence on Performance Management: The Case of a Danish Company in China

7. Thomas Nicolai Pedersen
The Discursive Constitution of Organizational Governance – Between unity and differentiation
The Case of the governance of environmental risks by World Bank environmental staff

8. Cynthia Selin
Volatile Visions: Transactons in Anticipatory Knowledge

9. Jesper Banghøj
Financial Accounting Information and Compensation in Danish Companies

10. Mikkel Lucas Overby
Strategic Alliances in Emerging High-Tech Markets: What's the Difference and does it Matter?

11. Tine Aage
External Information Acquisition of Industrial Districts and the Impact of Different Knowledge Creation Dimensions

A case study of the Fashion and Design Branch of the Industrial District of Montebelluna, NE Italy

12. Mikkel Flyverbom
Making the Global Information Society Governable
On the Governmentality of Multi-Stakeholder Networks

13. Anette Grønning
Personen bag
Tilstedeværer i e-mail som interaktionsform mellem kunde og medarbejder i dansk forsikringskontekst

14. Jørn Helder
One Company – One Language?
The NN-case

15. Lars Bjerregaard Mikkelsen
Differing perceptions of customer value
Development and application of a tool for mapping perceptions of customer value at both ends of customer-supplier dyads in industrial markets

16. Lise Granerud
Exploring Learning
Technological learning within small manufacturers in South Africa

17. Esben Rahbek Pedersen
Between Hopes and Realities: Reflections on the Promises and Practices of Corporate Social Responsibility (CSR)

18. Ramona Samson
The Cultural Integration Model and European Transformation.
The Case of Romania

2007

1. Jakob Vestergaard
Discipline in The Global Economy
Panopticism and the Post-Washington Consensus

2. Heidi Lund Hansen
Spaces for learning and working
A qualitative study of change of work, management, vehicles of power and social practices in open offices

3. Sudhanshu Rai
Exploring the internal dynamics of software development teams during user analysis
A tension enabled Institutionalization Model; "Where process becomes the objective"

4. Norsk ph.d.
Ej til salg gennem Samfundslitteratur

5. Serden Ozcan
EXPLORING HETEROGENEITY IN ORGANIZATIONAL ACTIONS AND OUTCOMES
A Behavioural Perspective

6. Kim Sundtoft Hald
Inter-organizational Performance Measurement and Management in Action
– An Ethnography on the Construction of Management, Identity and Relationships

7. Tobias Lindeberg
Evaluative Technologies
Quality and the Multiplicity of Performance

8. Merete Wedell-Wedellsborg
Den globale soldat
Identitetsdannelse og identitetsledelse i multinationale militære organisationer

9. Lars Frederiksen
Open Innovation Business Models
Innovation in firm-hosted online user communities and inter-firm project ventures in the music industry
– A collection of essays

10. Jonas Gabrielsen
Retorisk toposlære – fra statisk 'sted' til persuasiv aktivitet

11. Christian Moldt-Jørgensen
Fra meningsløs til meningsfuld evaluering.
Anvendelsen af studentertilfredshedsmålinger på de korte og mellemlange videregående uddannelser set fra et psykodynamisk systemperspektiv

12. Ping Gao
Extending the application of actor-network theory
Cases of innovation in the telecommunications industry

13. Peter Mejlby
Frihed og fængsel, en del af den samme drøm?
Et phronetisk baseret casestudie af frigørelsens og kontrollens sameksistens i værdibaseret ledelse!

14. Kristina Birch
Statistical Modelling in Marketing

15. Signe Poulsen
Sense and sensibility:
The language of emotional appeals in insurance marketing

16. Anders Bjerre Trolle
Essays on derivatives pricing and dynamic asset allocation

17. Peter Feldhütter
Empirical Studies of Bond and Credit Markets

18. Jens Henrik Eggert Christensen
Default and Recovery Risk Modeling and Estimation

19. Maria Theresa Larsen
Academic Enterprise: A New Mission for Universities or a Contradiction in Terms?
Four papers on the long-term implications of increasing industry involvement and commercialization in academia

20. Morten Wellendorf
Postimplementering af teknologi i den offentlige forvaltning
Analyser af en organisations kontinuerlige arbejde med informationsteknologi

21. Ekaterina Mhaanna
Concept Relations for Terminological Process Analysis

22. Stefan Ring Thorbjørnsen
Forsvaret i forandring
Et studie i officerers kapabiliteter under påvirkning af omverdenens forandringspres mod øget styring og læring

23. Christa Breum Amhøj
Det selvskabte medlemskab om managementstaten, dens styringsteknologier og indbyggere

24. Karoline Bromose
Between Technological Turbulence and Operational Stability
– An empirical case study of corporate venturing in TDC

25. Susanne Justesen
Navigating the Paradoxes of Diversity in Innovation Practice
– A Longitudinal study of six very different innovation processes – in practice

26. Luise Noring Henler
Conceptualising successful supply chain partnerships
– Viewing supply chain partnerships from an organisational culture perspective

27. Mark Mau
Kampen om telefonen
Det danske telefonvæsen under den tyske besættelse 1940-45

28. Jakob Halskov
The semiautomatic expansion of existing terminological ontologies using knowledge patterns discovered

on the WWW – an implementation and evaluation

29. Gergana Koleva
European Policy Instruments Beyond Networks and Structure: The Innovative Medicines Initiative

30. Christian Geisler Asmussen
Global Strategy and International Diversity: A Double-Edged Sword?

31. Christina Holm-Petersen
Stolthed og fordom
Kultur- og identitetsarbejde ved skabelsen af en ny sengeafdeling gennem fusion

32. Hans Peter Olsen
Hybrid Governance of Standardized States
Causes and Contours of the Global Regulation of Government Auditing

33. Lars Bøge Sørensen
Risk Management in the Supply Chain

34. Peter Aagaard
Det unikkes dynamikker
De institutionelle mulighedsbetingelser bag den individuelle udforskning i professionelt og frivilligt arbejde

35. Yun Mi Antorini
Brand Community Innovation
An Intrinsic Case Study of the Adult Fans of LEGO Community

36. Joachim Lynggaard Boll
Labor Related Corporate Social Performance in Denmark
Organizational and Institutional Perspectives

2008

1. Frederik Christian Vinten
Essays on Private Equity

2. Jesper Clement
Visual Influence of Packaging Design on In-Store Buying Decisions

3. Marius Brostrøm Kousgaard
Tid til kvalitetsmåling?
– Studier af indrulleringsprocesser i forbindelse med introduktionen af kliniske kvalitetsdatabaser i speciallægepraksissektoren

4. Irene Skovgaard Smith
Management Consulting in Action
Value creation and ambiguity in client-consultant relations

5. Anders Rom
Management accounting and integrated information systems
How to exploit the potential for management accounting of information technology

6. Marina Candi
Aesthetic Design as an Element of Service Innovation in New Technology-based Firms

7. Morten Schnack
Teknologi og tværfaglighed
– en analyse af diskussionen omkring indførelse af EPJ på en hospitalsafdeling

8. Helene Balslev Clausen
Juntos pero no revueltos – un estudio sobre emigrantes norteamericanos en un pueblo mexicano

9. Lise Justesen
Kunsten at skrive revisionsrapporter. En beretning om forvaltningsrevisionens beretninger

10. Michael E. Hansen
The politics of corporate responsibility: CSR and the governance of child labor and core labor rights in the 1990s

11. Anne Roepstorff
Holdning for handling – en etnologisk undersøgelse af Virksomheders Sociale Ansvar/CSR

12. Claus Bajlum
Essays on Credit Risk and Credit Derivatives

13. Anders Bojesen
The Performative Power of Competence – an Inquiry into Subjectivity and Social Technologies at Work

14. Satu Reijonen
Green and Fragile
A Study on Markets and the Natural Environment

15. Ilduara Busta
Corporate Governance in Banking
A European Study

16. Kristian Anders Hvass
A Boolean Analysis Predicting Industry Change: Innovation, Imitation & Business Models
The Winning Hybrid: A case study of isomorphism in the airline industry

17. Trine Paludan
De uvidende og de udviklingsparate
Identitet som mulighed og restriktion blandt fabriksarbejdere på det aftayloriserede fabriksgulv

18. Kristian Jakobsen
Foreign market entry in transition economies: Entry timing and mode choice

19. Jakob Elming
Syntactic reordering in statistical machine translation

20. Lars Brømsøe Termansen
Regional Computable General Equilibrium Models for Denmark
Three papers laying the foundation for regional CGE models with agglomeration characteristics

21. Mia Reinholt
The Motivational Foundations of Knowledge Sharing

22. Frederikke Krogh-Meibom
The Co-Evolution of Institutions and Technology
– A Neo-Institutional Understanding of Change Processes within the Business Press – the Case Study of Financial Times

23. Peter D. Ørberg Jensen
OFFSHORING OF ADVANCED AND HIGH-VALUE TECHNICAL SERVICES: ANTECEDENTS, PROCESS DYNAMICS AND FIRMLEVEL IMPACTS

24. Pham Thi Song Hanh
Functional Upgrading, Relational Capability and Export Performance of Vietnamese Wood Furniture Producers

25. Mads Vangkilde
Why wait?
An Exploration of first-mover advantages among Danish e-grocers through a resource perspective

26. Hubert Buch-Hansen
Rethinking the History of European Level Merger Control
A Critical Political Economy Perspective

2009

1. Vivian Lindhardsen
From Independent Ratings to Communal Ratings: A Study of CWA Raters' Decision-Making Behaviours

2. Guðrið Weihe
Public-Private Partnerships: Meaning and Practice

3. Chris Nøkkentved
Enabling Supply Networks with Collaborative Information Infrastructures
An Empirical Investigation of Business Model Innovation in Supplier Relationship Management

4. Sara Louise Muhr
Wound, Interrupted – On the Vulnerability of Diversity Management

5. Christine Sestoft
Forbrugeradfærd i et Stats- og Livsformsteoretisk perspektiv

6. Michael Pedersen
Tune in, Breakdown, and Reboot: On the production of the stress-fit self-managing employee

7. Salla Lutz
Position and Reposition in Networks – Exemplified by the Transformation of the Danish Pine Furniture Manufacturers

8. Jens Forssbæck
Essays on market discipline in commercial and central banking

9. Tine Murphy
Sense from Silence – A Basis for Organised Action
How do Sensemaking Processes with Minimal Sharing Relate to the Reproduction of Organised Action?

10. Sara Malou Strandvad
Inspirations for a new sociology of art: A sociomaterial study of development processes in the Danish film industry

11. Nicolaas Mouton
On the evolution of social scientific metaphors:
A cognitive-historical enquiry into the divergent trajectories of the idea that collective entities – states and societies, cities and corporations – are biological organisms.

12. Lars Andreas Knutsen
Mobile Data Services:
Shaping of user engagements

13. Nikolaos Theodoros Korfiatis
Information Exchange and Behavior
A Multi-method Inquiry on Online Communities

14. Jens Albæk
Forestillinger om kvalitet og tværfaglighed på sygehuse
– skabelse af forestillinger i læge- og plejegrupperne angående relevans af nye idéer om kvalitetsudvikling gennem tolkningsprocesser

15. Maja Lotz
The Business of Co-Creation – and the Co-Creation of Business

16. Gitte P. Jakobsen
Narrative Construction of Leader Identity in a Leader Development Program Context

17. Dorte Hermansen
"Living the brand" som en brandorienteret dialogisk praxis:
Om udvikling af medarbejdernes brandorienterede dømmekraft

18. Aseem Kinra
Supply Chain (logistics) Environmental Complexity

19. Michael Nørager
How to manage SMEs through the transformation from non innovative to innovative?

20. Kristin Wallevik
Corporate Governance in Family Firms
The Norwegian Maritime Sector

21. Bo Hansen Hansen
Beyond the Process
Enriching Software Process Improvement with Knowledge Management

22. Annemette Skot-Hansen
Franske adjektivisk afledte adverbier, der tager præpositionssyntagmer indledt med præpositionen à som argumenter
En valensgrammatisk undersøgelse

23. Line Gry Knudsen
Collaborative R&D Capabilities
In Search of Micro-Foundations

24. Christian Scheuer
Employers meet employees
Essays on sorting and globalization

25. Rasmus Johnsen
The Great Health of Melancholy
A Study of the Pathologies of Performativity

26. Ha Thi Van Pham
Internationalization, Competitiveness Enhancement and Export Performance of Emerging Market Firms:
Evidence from Vietnam

27. Henriette Balieu
Kontrolbegrebets betydning for kausativalternationen i spansk
En kognitiv-typologisk analyse

2010

1. Yen Tran
Organizing Innovationin Turbulent Fashion Market
Four papers on how fashion firms create and appropriate innovation value

2. Anders Raastrup Kristensen
Metaphysical Labour
Flexibility, Performance and Commitment in Work-Life Management

3. Margrét Sigrún Sigurdardottir
Dependently independent
Co-existence of institutional logics in the recorded music industry

4. Ásta Dis Óladóttir
Internationalization from a small domestic base:
An empirical analysis of Economics and Management

5. Christine Secher
E-deltagelse i praksis – politikernes og forvaltningens medkonstruktion og konsekvenserne heraf

6. Marianne Stang Våland
What we talk about when we talk about space:
End User Participation between Processes of Organizational and Architectural Design

7. Rex Degnegaard
Strategic Change Management
Change Management Challenges in the Danish Police Reform

8. Ulrik Schultz Brix
Værdi i rekruttering – den sikre beslutning
En pragmatisk analyse af perception og synliggørelse af værdi i rekrutterings- og udvælgelsesarbejdet

9. Jan Ole Similä
Kontraktsledelse
Relasjonen mellom virksomhetsledelse og kontraktshåndtering, belyst via fire norske virksomheter

10. Susanne Boch Waldorff
Emerging Organizations: In between local translation, institutional logics and discourse

11. Brian Kane
Performance Talk
Next Generation Management of Organizational Performance

12. Lars Ohnemus
Brand Thrust: Strategic Branding and Shareholder Value
An Empirical Reconciliation of two Critical Concepts

13. Jesper Schlamovitz
Håndtering af usikkerhed i film- og byggeprojekter

14. Tommy Moesby-Jensen
Det faktiske livs forbindtlighed
Førsokratisk informeret, ny-aristotelisk ἦθος*-tænkning hos Martin Heidegger*

15. Christian Fich
Two Nations Divided by Common Values
French National Habitus and the Rejection of American Power

16. Peter Beyer
Processer, sammenhængskraft og fleksibilitet
Et empirisk casestudie af omstillingsforløb i fire virksomheder

17. Adam Buchhorn
Markets of Good Intentions
Constructing and Organizing Biogas Markets Amid Fragility and Controversy

18. Cecilie K. Moesby-Jensen
Social læring og fælles praksis
Et mixed method studie, der belyser læringskonsekvenser af et lederkursus for et praksisfællesskab af offentlige mellemledere

19. Heidi Boye
Fødevarer og sundhed i senmodernismen
– En indsigt i hyggefænomenet og de relaterede fødevarepraksisser

20. Kristine Munkgård Pedersen
Flygtige forbindelser og midlertidige mobiliseringer
Om kulturel produktion på Roskilde Festival

21. Oliver Jacob Weber
Causes of Intercompany Harmony in Business Markets – An Empirical Investigation from a Dyad Perspective

22. Susanne Ekman
Authority and Autonomy
Paradoxes of Modern Knowledge Work

23. Anette Frey Larsen
Kvalitetsledelse på danske hospitaler
– Ledelsernes indflydelse på introduktion og vedligeholdelse af kvalitetsstrategier i det danske sundhedsvæsen

24. Toyoko Sato
Performativity and Discourse: Japanese Advertisements on the Aesthetic Education of Desire

25. Kenneth Brinch Jensen
Identifying the Last Planner System
Lean management in the construction industry

26. Javier Busquets
Orchestrating Network Behavior for Innovation

27. Luke Patey
The Power of Resistance: India's National Oil Company and International Activism in Sudan

28. Mette Vedel
Value Creation in Triadic Business Relationships. Interaction, Interconnection and Position

29. Kristian Tørning
Knowledge Management Systems in Practice – A Work Place Study

30. Qingxin Shi
An Empirical Study of Thinking Aloud Usability Testing from a Cultural Perspective

31. Tanja Juul Christiansen
Corporate blogging: Medarbejderes kommunikative handlekraft

32. Malgorzata Ciesielska
Hybrid Organisations.
A study of the Open Source – business setting

33. Jens Dick-Nielsen
Three Essays on Corporate Bond Market Liquidity

34. Sabrina Speiermann
Modstandens Politik
Kampagnestyring i Velfærdsstaten.
En diskussion af trafikkampagners styringspotentiale

35. Julie Uldam
Fickle Commitment. Fostering political engagement in 'the flighty world of online activism'

36. Annegrete Juul Nielsen
Traveling technologies and transformations in health care

37. Athur Mühlen-Schulte
Organising Development
Power and Organisational Reform in the United Nations Development Programme

38. Louise Rygaard Jonas
Branding på butiksgulvet
Et case-studie af kultur- og identitetsarbejdet i Kvickly

2011

1. Stefan Fraenkel
Key Success Factors for Sales Force Readiness during New Product Launch
A Study of Product Launches in the Swedish Pharmaceutical Industry

2. Christian Plesner Rossing
International Transfer Pricing in Theory and Practice

3. Tobias Dam Hede
Samtalekunst og ledelsesdisciplin
– en analyse af coachingsdiskursens genealogi og governmentality

4. Kim Pettersson
Essays on Audit Quality, Auditor Choice, and Equity Valuation

5. Henrik Merkelsen
The expert-lay controversy in risk research and management. Effects of institutional distances. Studies of risk definitions, perceptions, management and communication

6. Simon S. Torp
Employee Stock Ownership:
Effect on Strategic Management and Performance

7. Mie Harder
Internal Antecedents of Management Innovation

8. Ole Helby Petersen
Public-Private Partnerships: Policy and Regulation – With Comparative and Multi-level Case Studies from Denmark and Ireland

9. Morten Krogh Petersen
'Good' Outcomes. Handling Multiplicity in Government Communication

10. Kristian Tangsgaard Hvelplund
Allocation of cognitive resources in translation - an eye-tracking and key-logging study

11. Moshe Yonatany
The Internationalization Process of Digital Service Providers

12. Anne Vestergaard
Distance and Suffering
Humanitarian Discourse in the age of Mediatization

13. Thorsten Mikkelsen
Personligsheds indflydelse på forretningsrelationer

14. Jane Thostrup Jagd
Hvorfor fortsætter fusionsbølgen udover "the tipping point"?
– en empirisk analyse af information og kognitioner om fusioner

15. Gregory Gimpel
Value-driven Adoption and Consumption of Technology: Understanding Technology Decision Making

16. Thomas Stengade Sønderskov
Den nye mulighed
Social innovation i en forretningsmæssig kontekst

17. Jeppe Christoffersen
Donor supported strategic alliances in developing countries

18. Vibeke Vad Baunsgaard
Dominant Ideological Modes of Rationality: Cross functional

integration in the process of product innovation

19. Throstur Olaf Sigurjonsson
Governance Failure and Icelands's Financial Collapse

20. Allan Sall Tang Andersen
Essays on the modeling of risks in interest-rate and inflation markets

21. Heidi Tscherning
Mobile Devices in Social Contexts

22. Birgitte Gorm Hansen
Adapting in the Knowledge Economy Lateral Strategies for Scientists and Those Who Study Them

23. Kristina Vaarst Andersen
Optimal Levels of Embeddedness The Contingent Value of Networked Collaboration

24. Justine Grønbæk Pors
Noisy Management A History of Danish School Governing from 1970-2010

25. Stefan Linder
Micro-foundations of Strategic Entrepreneurship Essays on Autonomous Strategic Action

26. Xin Li
Toward an Integrative Framework of National Competitiveness An application to China

27. Rune Thorbjørn Clausen
Værdifuld arkitektur Et eksplorativt studie af bygningers rolle i virksomheders værdiskabelse

28. Monica Viken
Markedsundersøkelser som bevis i varemerke- og markedsføringsrett

29. Christian Wymann
Tattooing The Economic and Artistic Constitution of a Social Phenomenon

30. Sanne Frandsen
Productive Incoherence A Case Study of Branding and Identity Struggles in a Low-Prestige Organization

31. Mads Stenbo Nielsen
Essays on Correlation Modelling

32. Ivan Häuser
Følelse og sprog Etablering af en ekspressiv kategori, eksemplificeret på russisk

33. Sebastian Schwenen
Security of Supply in Electricity Markets

2012

1. Peter Holm Andreasen
The Dynamics of Procurement Management - A Complexity Approach

2. Martin Haulrich
Data-Driven Bitext Dependency Parsing and Alignment

3. Line Kirkegaard
Konsulenten i den anden nat En undersøgelse af det intense arbejdsliv

4. Tonny Stenheim
Decision usefulness of goodwill under IFRS

5. Morten Lind Larsen
Produktivitet, vækst og velfærd Industrirådet og efterkrigstidens Danmark 1945 - 1958

6. Petter Berg
Cartel Damages and Cost Asymmetries

7. Lynn Kahle
Experiential Discourse in Marketing A methodical inquiry into practice and theory

8. Anne Roelsgaard Obling
Management of Emotions in Accelerated Medical Relationships

9. Thomas Frandsen
Managing Modularity of Service Processes Architecture

10. Carina Christine Skovmøller
CSR som noget særligt
Et casestudie om styring og meningsskabelse i relation til CSR ud fra en intern optik

11. Michael Tell
Fradragsbeskæring af selskabers finansieringsudgifter
En skatteretlig analyse af SEL §§ 11, 11B og 11C

12. Morten Holm
Customer Profitability Measurement Models
Their Merits and Sophistication across Contexts

13. Katja Joo Dyppel
Beskatning af derivater
En analyse af dansk skatteret

14. Esben Anton Schultz
Essays in Labor Economics
Evidence from Danish Micro Data

15. Carina Risvig Hansen
"Contracts not covered, or not fully covered, by the Public Sector Directive"

16. Anja Svejgaard Pors
Iværksættelse af kommunikation
- patientfigurer i hospitalets strategiske kommunikation

17. Frans Bévort
Making sense of management with logics
An ethnographic study of accountants who become managers

18. René Kallestrup
The Dynamics of Bank and Sovereign Credit Risk

19. Brett Crawford
Revisiting the Phenomenon of Interests in Organizational Institutionalism
The Case of U.S. Chambers of Commerce

20. Mario Daniele Amore
Essays on Empirical Corporate Finance

21. Arne Stjernholm Madsen
The evolution of innovation strategy
Studied in the context of medical device activities at the pharmaceutical company Novo Nordisk A/S in the period 1980-2008

22. Jacob Holm Hansen
Is Social Integration Necessary for Corporate Branding?
A study of corporate branding strategies at Novo Nordisk

23. Stuart Webber
Corporate Profit Shifting and the Multinational Enterprise

24. Helene Ratner
Promises of Reflexivity
Managing and Researching Inclusive Schools

25. Therese Strand
The Owners and the Power: Insights from Annual General Meetings

26. Robert Gavin Strand
In Praise of Corporate Social Responsibility Bureaucracy

27. Nina Sormunen
Auditor's going-concern reporting
Reporting decision and content of the report

28. John Bang Mathiasen
Learning within a product development working practice:
- an understanding anchored in pragmatism

29. Philip Holst Riis
Understanding Role-Oriented Enterprise Systems: From Vendors to Customers

30. Marie Lisa Dacanay
Social Enterprises and the Poor
Enhancing Social Entrepreneurship and Stakeholder Theory

31. Fumiko Kano Glückstad
Bridging Remote Cultures: Cross-lingual concept mapping based on the information receiver's prior-knowledge

32. Henrik Barslund Fosse
Empirical Essays in International Trade

33. Peter Alexander Albrecht
Foundational hybridity and its reproduction
Security sector reform in Sierra Leone

34. Maja Rosenstock
CSR - hvor svært kan det være? Kulturanalytisk casestudie om udfordringer og dilemmaer med at forankre Coops CSR-strategi

35. Jeanette Rasmussen
Tweens, medier og forbrug
Et studie af 10-12 årige danske børns brug af internettet, opfattelse og forståelse af markedsføring og forbrug

36. Ib Tunby Gulbrandsen
'This page is not intended for a US Audience'
A five-act spectacle on online communication, collaboration & organization.

37. Kasper Aalling Teilmann
Interactive Approaches to Rural Development

38. Mette Mogensen
The Organization(s) of Well-being and Productivity
(Re)assembling work in the Danish Post

39. Søren Friis Møller
From Disinterestedness to Engagement
Towards Relational Leadership In the Cultural Sector

40. Nico Peter Berhausen
Management Control, Innovation and Strategic Objectives – Interactions and Convergence in Product Development Networks

41. Balder Onarheim
Creativity under Constraints
Creativity as Balancing 'Constrainedness'

42. Haoyong Zhou
Essays on Family Firms

43. Elisabeth Naima Mikkelsen
Making sense of organisational conflict
An empirical study of enacted sensemaking in everyday conflict at work

2013

1. Jacob Lyngsie
Entrepreneurship in an Organizational Context

2. Signe Groth-Brodersen
Fra ledelse til selvet
En socialpsykologisk analyse af forholdet imellem selvledelse, ledelse og stress i det moderne arbejdsliv

3. Nis Høyrup Christensen
Shaping Markets: A Neoinstitutional Analysis of the Emerging Organizational Field of Renewable Energy in China

4. Christian Edelvold Berg
As a matter of size
THE IMPORTANCE OF CRITICAL MASS AND THE CONSEQUENCES OF SCARCITY FOR TELEVISION MARKETS

5. Christine D. Isakson
Coworker Influence and Labor Mobility
Essays on Turnover, Entrepreneurship and Location Choice in the Danish Maritime Industry

6. Niels Joseph Jerne Lennon
Accounting Qualities in Practice
Rhizomatic stories of representational faithfulness, decision making and control

7. Shannon O'Donnell
Making Ensemble Possible
How special groups organize for collaborative creativity in conditions of spatial variability and distance

8. Robert W. D. Veitch
Access Decisions in a Partly-Digital World
Comparing Digital Piracy and Legal Modes for Film and Music

9. Marie Mathiesen
Making Strategy Work
An Organizational Ethnography

10. Arisa Shollo
The role of business intelligence in organizational decision-making

11. Mia Kaspersen
The construction of social and environmental reporting

12. Marcus Møller Larsen
The organizational design of offshoring

13. Mette Ohm Rørdam
EU Law on Food Naming
The prohibition against misleading names in an internal market context

14. Hans Peter Rasmussen
GIV EN GED!
Kan giver-idealtyper forklare støtte til velgørenhed og understøtte relationsopbygning?

15. Ruben Schachtenhaufen
Fonetisk reduktion i dansk

16. Peter Koerver Schmidt
Dansk CFC-beskatning
I et internationalt og komparativt perspektiv

17. Morten Froholdt
Strategi i den offentlige sektor
En kortlægning af styringsmæssig kontekst, strategisk tilgang, samt anvendte redskaber og teknologier for udvalgte danske statslige styrelser

18. Annette Camilla Sjørup
Cognitive effort in metaphor translation
An eye-tracking and key-logging study

19. Tamara Stucchi
The Internationalization of Emerging Market Firms:
A Context-Specific Study

20. Thomas Lopdrup-Hjorth
"Let's Go Outside":
The Value of Co-Creation

21. Ana Alačovska
Genre and Autonomy in Cultural Production
The case of travel guidebook production

22. Marius Gudmand-Høyer
Stemningssindssygdommenes historie i det 19. århundrede
Omtydningen af melankolien og manien som bipolære stemningslidelser i dansk sammenhæng under hensyn til dannelsen af det moderne følelseslivs relative autonomi.
En problematiserings- og erfarings-analytisk undersøgelse

23. Lichen Alex Yu
Fabricating an S&OP Process
Circulating References and Matters of Concern

24. Esben Alfort
The Expression of a Need
Understanding search

25. Trine Pallesen
Assembling Markets for Wind Power
An Inquiry into the Making of Market Devices

26. Anders Koed Madsen
Web-Visions
Repurposing digital traces to organize social attention

27. Lærke Højgaard Christiansen
BREWING ORGANIZATIONAL RESPONSES TO INSTITUTIONAL LOGICS

28. Tommy Kjær Lassen
EGENTLIG SELVLEDELSE
En ledelsesfilosofisk afhandling om selvledelsens paradoksale dynamik og eksistentielle engagement

29. Morten Rossing
Local Adaption and Meaning Creation in Performance Appraisal

30. Søren Obed Madsen
Lederen som oversætter
Et oversættelsesteoretisk perspektiv på strategisk arbejde

31. Thomas Høgenhaven
Open Government Communities
Does Design Affect Participation?

32. Kirstine Zinck Pedersen
Failsafe Organizing?
A Pragmatic Stance on Patient Safety

33. Anne Petersen
Hverdagslogikker i psykiatrisk arbejde
En institutionsetnografisk undersøgelse af hverdagen i psykiatriske organisationer

34. Didde Maria Humle
Fortællinger om arbejde

35. Mark Holst-Mikkelsen
Strategieksekvering i praksis
– barrierer og muligheder!

36. Malek Maalouf
Sustaining lean
Strategies for dealing with organizational paradoxes

37. Nicolaj Tofte Brenneche
Systemic Innovation In The Making
The Social Productivity of Cartographic Crisis and Transitions in the Case of SEEIT

38. Morten Gylling
The Structure of Discourse
A Corpus-Based Cross-Linguistic Study

39. Binzhang YANG
Urban Green Spaces for Quality Life
- Case Study: the landscape architecture for people in Copenhagen

40. Michael Friis Pedersen
Finance and Organization:
The Implications for Whole Farm Risk Management

41. Even Fallan
Issues on supply and demand for environmental accounting information

42. Ather Nawaz
Website user experience
A cross-cultural study of the relation between users´ cognitive style, context of use, and information architecture of local websites

43. Karin Beukel
The Determinants for Creating Valuable Inventions

44. Arjan Markus
External Knowledge Sourcing and Firm Innovation
Essays on the Micro-Foundations of Firms' Search for Innovation

2014

1. Solon Moreira
Four Essays on Technology Licensing and Firm Innovation

2. Karin Strzeletz Ivertsen
Partnership Drift in Innovation Processes
A study of the Think City electric car development

3. Kathrine Hoffmann Pii
Responsibility Flows in Patient-centred Prevention

4. Jane Bjørn Vedel
Managing Strategic Research
An empirical analysis of science-industry collaboration in a pharmaceutical company

5. Martin Gylling
Processuel strategi i organisationer
Monografi om dobbeltheden i tænkning af strategi, dels som vidensfelt i organisationsteori, dels som kunstnerisk tilgang til at skabe i erhvervsmæssig innovation

6. Linne Marie Lauesen
Corporate Social Responsibility in the Water Sector: How Material Practices and their Symbolic and Physical Meanings Form a Colonising Logic

7. Maggie Qiuzhu Mei
LEARNING TO INNOVATE: The role of ambidexterity, standard, and decision process

8. Inger Høedt-Rasmussen
Developing Identity for Lawyers Towards Sustainable Lawyering

9. Sebastian Fux
Essays on Return Predictability and Term Structure Modelling

10. Thorbjørn N. M. Lund-Poulsen
Essays on Value Based Management

11. Oana Brindusa Albu
Transparency in Organizing: A Performative Approach

12. Lena Olaison
Entrepreneurship at the limits

13. Hanne Sørum
DRESSED FOR WEB SUCCESS? An Empirical Study of Website Quality in the Public Sector

14. Lasse Folke Henriksen
Knowing networks How experts shape transnational governance

15. Maria Halbinger
Entrepreneurial Individuals Empirical Investigations into Entrepreneurial Activities of Hackers and Makers

16. Robert Spliid
Kapitalfondenes metoder og kompetencer

17. Christiane Stelling
Public-private partnerships & the need, development and management of trusting A processual and embedded exploration

18. Marta Gasparin
Management of design as a translation process

19. Kåre Moberg
Assessing the Impact of Entrepreneurship Education From ABC to PhD

20. Alexander Cole
Distant neighbors Collective learning beyond the cluster

21. Martin Møller Boje Rasmussen
Is Competitiveness a Question of Being Alike? How the United Kingdom, Germany and Denmark Came to Compete through their Knowledge Regimes from 1993 to 2007

22. Anders Ravn Sørensen
Studies in central bank legitimacy, currency and national identity Four cases from Danish monetary history

23. Nina Bellak
Can Language be Managed in International Business? Insights into Language Choice from a Case Study of Danish and Austrian Multinational Corporations (MNCs)

24. Rikke Kristine Nielsen
Global Mindset as Managerial Meta-competence and Organizational Capability: Boundary-crossing Leadership Cooperation in the MNC The Case of 'Group Mindset' in Solar A/S.

25. Rasmus Koss Hartmann
User Innovation inside government Towards a critically performative foundation for inquiry

26. Kristian Gylling Olesen
Flertydig og emergerende ledelse i folkeskolen
Et aktør-netværksteoretisk ledelses-studie af politiske evalueringsreformers betydning for ledelse i den danske folkeskole

27. Troels Riis Larsen
Kampen om Danmarks omdømme 1945-2010
Omdømmearbejde og omdømmepolitik

28. Klaus Majgaard
Jagten på autenticitet i offentlig styring

29. Ming Hua Li
Institutional Transition and Organizational Diversity:
Differentiated internationalization strategies of emerging market state-owned enterprises

30. Sofie Blinkenberg Federspiel
IT, organisation og digitalisering:
Institutionelt arbejde i den kommunale digitaliseringsproces

31. Elvi Weinreich
Hvilke offentlige ledere er der brug for når velfærdstænkningen flytter sig – er Diplomuddannelsens lederprofil svaret?

32. Ellen Mølgaard Korsager
Self-conception and image of context in the growth of the firm – A Penrosian History of Fiberline Composites

33. Else Skjold
The Daily Selection

34. Marie Louise Conradsen
The Cancer Centre That Never Was
The Organisation of Danish Cancer Research 1949-1992

35. Virgilio Failla
Three Essays on the Dynamics of Entrepreneurs in the Labor Market

36. Nicky Nedergaard
Brand-Based Innovation
Relational Perspectives on Brand Logics and Design Innovation Strategies and Implementation

37. Mads Gjedsted Nielsen
Essays in Real Estate Finance

38. Kristin Martina Brandl
Process Perspectives on Service Offshoring

39. Mia Rosa Koss Hartmann
In the gray zone
With police in making space for creativity

40. Karen Ingerslev
Healthcare Innovation under The Microscope
Framing Boundaries of Wicked Problems

41. Tim Neerup Themsen
Risk Management in large Danish public capital investment programmes

2015

1. Jakob Ion Wille
Film som design
Design af levende billeder i film og tv-serier

2. Christiane Mossin
Interzones of Law and Metaphysics
Hierarchies, Logics and Foundations of Social Order seen through the Prism of EU Social Rights

3. Thomas Tøth
TRUSTWORTHINESS: ENABLING GLOBAL COLLABORATION
An Ethnographic Study of Trust, Distance, Control, Culture and Boundary Spanning within Offshore Outsourcing of IT Services

4. Steven Højlund
Evaluation Use in Evaluation Systems – The Case of the European Commission

5. Julia Kirch Kirkegaard
AMBIGUOUS WINDS OF CHANGE – OR FIGHTING AGAINST WINDMILLS IN CHINESE WIND POWER
A CONSTRUCTIVIST INQUIRY INTO CHINA'S PRAGMATICS OF GREEN MARKETISATION MAPPING CONTROVERSIES OVER A POTENTIAL TURN TO QUALITY IN CHINESE WIND POWER

6. Michelle Carol Antero
A Multi-case Analysis of the Development of Enterprise Resource Planning Systems (ERP) Business Practices

Morten Friis-Olivarius
The Associative Nature of Creativity

7. Mathew Abraham
New Cooperativism:
A study of emerging producer organisations in India

8. Stine Hedegaard
Sustainability-Focused Identity: Identity work performed to manage, negotiate and resolve barriers and tensions that arise in the process of constructing or ganizational identity in a sustainability context

9. Cecilie Glerup
Organizing Science in Society – the conduct and justification of resposible research

10. Allan Salling Pedersen
Implementering af ITIL® IT-governance - når best practice konflikter med kulturen Løsning af implementerings-problemer gennem anvendelse af kendte CSF i et aktionsforskningsforløb.

11. Nihat Misir
A Real Options Approach to Determining Power Prices

12. Mamdouh Medhat
MEASURING AND PRICING THE RISK OF CORPORATE FAILURES

13. Rina Hansen
Toward a Digital Strategy for Omnichannel Retailing

14. Eva Pallesen
In the rhythm of welfare creation
A relational processual investigation moving beyond the conceptual horizon of welfare management

15. Gouya Harirchi
In Search of Opportunities: Three Essays on Global Linkages for Innovation

16. Lotte Holck
Embedded Diversity: A critical ethnographic study of the structural tensions of organizing diversity

17. Jose Daniel Balarezo
Learning through Scenario Planning

18. Louise Pram Nielsen
Knowledge dissemination based on terminological ontologies. Using eye tracking to further user interface design.

19. Sofie Dam
PUBLIC-PRIVATE PARTNERSHIPS FOR INNOVATION AND SUSTAINABILITY TRANSFORMATION
An embedded, comparative case study of municipal waste management in England and Denmark

20. Ulrik Hartmyer Christiansen
Follwoing the Content of Reported Risk Across the Organization

21. Guro Refsum Sanden
Language strategies in multinational corporations. A cross-sector study of financial service companies and manufacturing companies.

22. Linn Gevoll
Designing performance management for operational level
- A closer look on the role of design choices in framing coordination and motivation

23. Frederik Larsen
Objects and Social Actions – on Second-hand Valuation Practices

24. Thorhildur Hansdottir Jetzek
The Sustainable Value of Open Government Data
Uncovering the Generative Mechanisms of Open Data through a Mixed Methods Approach

25. Gustav Toppenberg
Innovation-based M&A – Technological-Integration Challenges – The Case of Digital-Technology Companies

26. Mie Plotnikof
Challenges of Collaborative Governance
An Organizational Discourse Study of Public Managers' Struggles with Collaboration across the Daycare Area

27. Christian Garmann Johnsen
Who Are the Post-Bureaucrats? A Philosophical Examination of the Creative Manager, the Authentic Leader and the Entrepreneur

28. Jacob Brogaard-Kay
Constituting Performance Management A field study of a pharmaceutical company

29. Rasmus Ploug Jenle
Engineering Markets for Control: Integrating Wind Power into the Danish Electricity System

30. Morten Lindholst
Complex Business Negotiation: Understanding Preparation and Planning

31. Morten Grynings
TRUST AND TRANSPARENCY FROM AN ALIGNMENT PERSPECTIVE

32. Peter Andreas Norn
Byregimer og styringsevne: Politisk lederskab af store byudviklingsprojekter

33. Milan Miric
Essays on Competition, Innovation and Firm Strategy in Digital Markets

34. Sanne K. Hjordrup
The Value of Talent Management Rethinking practice, problems and possibilities

35. Johanna Sax
Strategic Risk Management – Analyzing Antecedents and Contingencies for Value Creation

36. Pernille Rydén
Strategic Cognition of Social Media

37. Mimmi Sjöklint
The Measurable Me - The Influence of Self-tracking on the User Experience

38. Juan Ignacio Staricco
Towards a Fair Global Economic Regime? A critical assessment of Fair Trade through the examination of the Argentinean wine industry

39. Marie Henriette Madsen
Emerging and temporary connections in Quality work

40. Yangfeng CAO
Toward a Process Framework of Business Model Innovation in the Global Context
Entrepreneurship-Enabled Dynamic Capability of Medium-Sized Multinational Enterprises

41. Carsten Scheibye
Enactment of the Organizational Cost Structure in Value Chain Configuration A Contribution to Strategic Cost Management

2016

1. Signe Sofie Dyrby
Enterprise Social Media at Work

2. Dorte Boesby Dahl
The making of the public parking attendant
Dirt, aesthetics and inclusion in public service work

3. Verena Girschik
Realizing Corporate Responsibility Positioning and Framing in Nascent Institutional Change

4. Anders Ørding Olsen
IN SEARCH OF SOLUTIONS
Inertia, Knowledge Sources and Diversity in Collaborative Problem-solving

5. Pernille Steen Pedersen
Udkast til et nyt copingbegreb
En kvalifikation af ledelsesmuligheder for at forebygge sygefravær ved psykiske problemer.

6. Kerli Kant Hvass
Weaving a Path from Waste to Value: Exploring fashion industry business models and the circular economy

7. Kasper Lindskow
Exploring Digital News Publishing Business Models – a production network approach

8. Mikkel Mouritz Marfelt
The chameleon workforce: Assembling and negotiating the content of a workforce

9. Marianne Bertelsen
Aesthetic encounters
Rethinking autonomy, space & time in today's world of art

10. Louise Hauberg Wilhelmsen
EU PERSPECTIVES ON INTERNATIONAL COMMERCIAL ARBITRATION

11. Abid Hussain
On the Design, Development and Use of the Social Data Analytics Tool (SODATO): Design Propositions, Patterns, and Principles for Big Social Data Analytics

12. Mark Bruun
Essays on Earnings Predictability

13. Tor Bøe-Lillegraven
BUSINESS PARADOXES, BLACK BOXES, AND BIG DATA: BEYOND ORGANIZATIONAL AMBIDEXTERITY

14. Hadis Khonsary-Atighi
ECONOMIC DETERMINANTS OF DOMESTIC INVESTMENT IN AN OIL-BASED ECONOMY: THE CASE OF IRAN (1965-2010)

15. Maj Lervad Grasten
Rule of Law or Rule by Lawyers?
On the Politics of Translation in Global Governance

16. Lene Granzau Juel-Jacobsen
SUPERMARKEDETS MODUS OPERANDI – en hverdagssociologisk undersøgelse af forholdet mellem rum og handlen og understøtte relationsopbygning?

17. Christine Thalsgård Henriques
In search of entrepreneurial learning – Towards a relational perspective on incubating practices?

18. Patrick Bennett
Essays in Education, Crime, and Job Displacement

19. Søren Korsgaard
Payments and Central Bank Policy

20. Marie Kruse Skibsted
Empirical Essays in Economics of Education and Labor

21. Elizabeth Benedict Christensen
The Constantly Contingent Sense of Belonging of the 1.5 Generation Undocumented Youth
An Everyday Perspective

22. Lasse J. Jessen
Essays on Discounting Behavior and Gambling Behavior

23. Kalle Johannes Rose
Når stifterviljen dør…
Et retsøkonomisk bidrag til 200 års juridisk konflikt om ejendomsretten

24. Andreas Søeborg Kirkedal
Danish Stød and Automatic Speech Recognition

25. Ida Lunde Jørgensen
Institutions and Legitimations in Finance for the Arts

26. Olga Rykov Ibsen
An empirical cross-linguistic study of directives: A semiotic approach to the sentence forms chosen by British, Danish and Russian speakers in native and ELF contexts

27. Desi Volker
Understanding Interest Rate Volatility

28. Angeli Elizabeth Weller
Practice at the Boundaries of Business Ethics & Corporate Social Responsibility

29. Ida Danneskiold-Samsøe
Levende læring i kunstneriske organisationer
En undersøgelse af læringsprocesser mellem projekt og organisation på Aarhus Teater

30. Leif Christensen
Quality of information – The role of internal controls and materiality

31. Olga Zarzecka
Tie Content in Professional Networks

32. Henrik Mahncke
De store gaver
- Filantropiens gensidighedsrelationer i teori og praksis

33. Carsten Lund Pedersen
Using the Collective Wisdom of Frontline Employees in Strategic Issue Management

34. Yun Liu
Essays on Market Design

35. Denitsa Hazarbassanova Blagoeva
The Internationalisation of Service Firms

36. Manya Jaura Lind
Capability development in an off-shoring context: How, why and by whom

37. Luis R. Boscán F.
Essays on the Design of Contracts and Markets for Power System Flexibility

38. Andreas Philipp Distel
Capabilities for Strategic Adaptation: Micro-Foundations, Organizational Conditions, and Performance Implications

39. Lavinia Bleoca
The Usefulness of Innovation and Intellectual Capital in Business Performance: The Financial Effects of Knowledge Management vs. Disclosure

40. Henrik Jensen
Economic Organization and Imperfect Managerial Knowledge: A Study of the Role of Managerial Meta-Knowledge in the Management of Distributed Knowledge

41. Stine Mosekjær
The Understanding of English Emotion Words by Chinese and Japanese Speakers of English as a Lingua Franca An Empirical Study

42. Hallur Tor Sigurdarson
The Ministry of Desire - Anxiety and entrepreneurship in a bureaucracy

43. Kätlin Pulk
Making Time While Being in Time
A study of the temporality of organizational processes

44. Valeria Giacomin
Contextualizing the cluster Palm oil in Southeast Asia in global perspective (1880s–1970s)

45. Jeanette Willert
Managers' use of multiple Management Control Systems: The role and interplay of management control systems and company performance

46. Mads Vestergaard Jensen
Financial Frictions: Implications for Early Option Exercise and Realized Volatility

47. Mikael Reimer Jensen
Interbank Markets and Frictions

48. Benjamin Faigen
Essays on Employee Ownership

49. Adela Michea
Enacting Business Models
An Ethnographic Study of an Emerging Business Model Innovation within the Frame of a Manufacturing Company.

50. Iben Sandal Stjerne
Transcending organization in temporary systems
Aesthetics' organizing work and employment in Creative Industries

51. Simon Krogh
Anticipating Organizational Change

52. Sarah Netter
Exploring the Sharing Economy

53. Lene Tolstrup Christensen
State-owned enterprises as institutional market actors in the marketization of public service provision:
A comparative case study of Danish and Swedish passenger rail 1990–2015

54. Kyoung(Kay) Sun Park
Three Essays on Financial Economics

2017

1. Mari Bjerck
Apparel at work. Work uniforms and women in male-dominated manual occupations.

2. Christoph H. Flöthmann
Who Manages Our Supply Chains? Backgrounds, Competencies and Contributions of Human Resources in Supply Chain Management

3. Aleksandra Anna Rzeźnik
Essays in Empirical Asset Pricing

4. Claes Bäckman
Essays on Housing Markets

5. Kirsti Reitan Andersen
Stabilizing Sustainability in the Textile and Fashion Industry

6. Kira Hoffmann
Cost Behavior: An Empirical Analysis of Determinants and Consequences of Asymmetries

7. Tobin Hanspal
Essays in Household Finance

8. Nina Lange
Correlation in Energy Markets

9. Anjum Fayyaz
Donor Interventions and SME Networking in Industrial Clusters in Punjab Province, Pakistan

10. Magnus Paulsen Hansen
Trying the unemployed. Justification and critique, emancipation and coercion towards the 'active society'. A study of contemporary reforms in France and Denmark

11. Sameer Azizi
Corporate Social Responsibility in Afghanistan
– a critical case study of the mobile telecommunications industry

12. Malene Myhre
The internationalization of small and medium-sized enterprises: A qualitative study

13. Thomas Presskorn-Thygesen
The Significance of Normativity – Studies in Post-Kantian Philosophy and Social Theory

14. Federico Clementi
Essays on multinational production and international trade

15. Lara Anne Hale
Experimental Standards in Sustainability Transitions: Insights from the Building Sector

16. Richard Pucci
Accounting for Financial Instruments in an Uncertain World Controversies in IFRS in the Aftermath of the 2008 Financial Crisis

17. Sarah Maria Denta
Kommunale offentlige private partnerskaber Regulering I skyggen af Farumsagen

18. Christian Östlund
Design for e-training

19. Amalie Martinus Hauge
Organizing Valuations – a pragmatic inquiry

20. Tim Holst Celik
Tension-filled Governance? Exploring the Emergence, Consolidation and Reconfiguration of Legitimatory and Fiscal State-crafting

21. Christian Bason
Leading Public Design: How managers engage with design to transform public governance

22. Davide Tomio
Essays on Arbitrage and Market Liquidity

23. Simone Stæhr
Financial Analysts' Forecasts Behavioral Aspects and the Impact of Personal Characteristics

24. Mikkel Godt Gregersen
Management Control, Intrinsic Motivation and Creativity – How Can They Coexist

25. Kristjan Johannes Suse Jespersen
Advancing the Payments for Ecosystem Service Discourse Through Institutional Theory

26. Kristian Bondo Hansen
Crowds and Speculation: A study of crowd phenomena in the U.S. financial markets 1890 to 1940

27. Lars Balslev
Actors and practices – An institutional study on management accounting change in Air Greenland

28. Sven Klingler
Essays on Asset Pricing with Financial Frictions

29. Klement Ahrensbach Rasmussen
Business Model Innovation The Role of Organizational Design

30. Giulio Zichella
Entrepreneurial Cognition. Three essays on entrepreneurial behavior and cognition under risk and uncertainty

31. Richard Ledborg Hansen
En forkærlighed til det eksisterende – mellemlederens oplevelse af forandringsmodstand i organisatoriske forandringer

32. Vilhelm Stefan Holsting
Militært chefvirke: Kritik og retfærdiggørelse mellem politik og profession

33. Thomas Jensen
Shipping Information Pipeline: An information infrastructure to improve international containerized shipping

34. Dzmitry Bartalevich
Do economic theories inform policy? Analysis of the influence of the Chicago School on European Union competition policy

35. Kristian Roed Nielsen
Crowdfunding for Sustainability: A study on the potential of reward-based crowdfunding in supporting sustainable entrepreneurship

36. Emil Husted
There is always an alternative: A study of control and commitment in political organization

37. Anders Ludvig Sevelsted
Interpreting Bonds and Boundaries of Obligation. A genealogy of the emergence and development of Protestant voluntary social work in Denmark as shown through the cases of the Copenhagen Home Mission and the Blue Cross (1850 – 1950)

38. Niklas Kohl
Essays on Stock Issuance

39. Maya Christiane Flensborg Jensen
BOUNDARIES OF PROFESSIONALIZATION AT WORK An ethnography-inspired study of care workers' dilemmas at the margin

40. Andreas Kamstrup
Crowdsourcing and the Architectural Competition as Organisational Technologies

41. Louise Lyngfeldt Gorm Hansen
Triggering Earthquakes in Science, Politics and Chinese Hydropower - A Controversy Study

2018

1. Vishv Priya Kohli
Combatting Falsifi cation and Counterfeiting of Medicinal Products in the E uropean Union – A Legal Analysis

2. Helle Haurum
Customer Engagement Behavior in the context of Continuous Service Relationships

3. Nis Grünberg
The Party -state order: Essays on China's political organization and political economic institutions

4. Jesper Christensen
A Behavioral Theory of Human Capital Integration

5. Poula Marie Helth
Learning in practice

6. Rasmus Vendler Toft-Kehler
Entrepreneurship as a career? An investigation of the relationship between entrepreneurial experience and entrepreneurial outcome

7. Szymon Furtak
Sensing the Future: Designing sensor-based predictive information systems for forecasting spare part demand for diesel engines

8. Mette Brehm Johansen
Organizing patient involvement. An ethnographic study

9. Iwona Sulinska
Complexities of Social Capital in Boards of Directors

10. Cecilie Fanøe Petersen
Award of public contracts as a means to conferring State aid: A legal analysis of the interface between public procurement law and State aid law

11. Ahmad Ahmad Barirani
Three Experimental Studies on Entrepreneurship

12. Carsten Allerslev Olsen
Financial Reporting Enforcement: Impact and Consequences

13. Irene Christensen
New product fumbles – Organizing for the Ramp-up process

14. Jacob Taarup-Esbensen
Managing communities – Mining MNEs' community risk management practices

15. Lester Allan Lasrado
Set-Theoretic approach to maturity models

16. Mia B. Münster
Intention vs. Perception of Designed Atmospheres in Fashion Stores

17. Anne Sluhan
Non-Financial Dimensions of Family Firm Ownership: How Socioemotional Wealth and Familiness Influence Internationalization

18. Henrik Yde Andersen
Essays on Debt and Pensions

19. Fabian Heinrich Müller
Valuation Reversed – When Valuators are Valuated. An Analysis of the Perception of and Reaction to Reviewers in Fine-Dining

20. Martin Jarmatz
Organizing for Pricing

21. Niels Joachim Christfort Gormsen
Essays on Empirical Asset Pricing

22. Diego Zunino
Socio-Cognitive Perspectives in Business Venturing

23. Benjamin Asmussen
Networks and Faces between Copenhagen and Canton, 1730-1840

24. Dalia Bagdziunaite
Brains at Brand Touchpoints A Consumer Neuroscience Study of Information Processing of Brand Advertisements and the Store Environment in Compulsive Buying

25. Erol Kazan
Towards a Disruptive Digital Platform Model

26. Andreas Bang Nielsen
Essays on Foreign Exchange and Credit Risk

27. Anne Krebs
Accountable, Operable Knowledge Toward Value Representations of Individual Knowledge in Accounting

28. Matilde Fogh Kirkegaard
A firm- and demand-side perspective on behavioral strategy for value creation: Insights from the hearing aid industry

29. Agnieszka Nowinska
SHIPS AND RELATION-SHIPS Tie formation in the sector of shipping intermediaries in shipping

30. Stine Evald Bentsen
The Comprehension of English Texts by Native Speakers of English and Japanese, Chinese and Russian Speakers of English as a Lingua Franca. An Empirical Study.

31. Stine Louise Daetz
Essays on Financial Frictions in Lending Markets

32. *Christian Skov Jensen*
Essays on Asset Pricing

33. Anders Kryger
Aligning future employee action and corporate strategy in a resource-scarce environment

34. Maitane Elorriaga-Rubio
The behavioral foundations of strategic decision-making: A contextual perspective

35. Roddy Walker
Leadership Development as Organisational Rehabilitation: Shaping Middle-Managers as Double Agents

36. Jinsun Bae
Producing Garments for Global Markets Corporate social responsibility (CSR) in Myanmar's export garment industry 2011–2015

37. Queralt Prat-i-Pubill
Axiological knowledge in a knowledge driven world. Considerations for organizations.

38. Pia Mølgaard
Essays on Corporate Loans and Credit Risk

39. Marzia Aricò
Service Design as a Transformative Force: Introduction and Adoption in an Organizational Context

40. Christian Dyrlund Wåhlin-Jacobsen
Constructing change initiatives in workplace voice activities Studies from a social interaction perspective

41. Peter Kalum Schou
Institutional Logics in Entrepreneurial Ventures: How Competing Logics arise and shape organizational processes and outcomes during scale-up

42. Per Henriksen
Enterprise Risk Management Rationaler og paradokser i en moderne ledelsesteknologi

43. Maximilian Schellmann
The Politics of Organizing Refugee Camps

44. Jacob Halvas Bjerre
Excluding the Jews: The Aryanization of Danish-German Trade and German Anti-Jewish Policy in Denmark 1937-1943

45. Ida Schrøder
Hybridising accounting and caring: A symmetrical study of how costs and needs are connected in Danish child protection work

46. Katrine Kunst
Electronic Word of Behavior: Transforming digital traces of consumer behaviors into communicative content in product design

47. Viktor Avlonitis
Essays on the role of modularity in management: Towards a unified perspective of modular and integral design

48. Anne Sofie Fischer
Negotiating Spaces of Everyday Politics: -An ethnographic study of organizing for social transformation for women in urban poverty, Delhi, India

2019

1. Shihan Du
   ESSAYS IN EMPIRICAL STUDIES BASED ON ADMINISTRATIVE LABOUR MARKET DATA

2. Mart Laatsit
   Policy learning in innovation policy: A comparative analysis of European Union member states

3. Peter J. Wynne
   Proactively Building Capabilities for the Post-Acquisition Integration of Information Systems

4. Kalina S. Staykova
   Generative Mechanisms for Digital Platform Ecosystem Evolution

5. Ieva Linkeviciute
   Essays on the Demand-Side Management in Electricity Markets

6. Jonatan Echebarria Fernández
   Jurisdiction and Arbitration Agreements in Contracts for the Carriage of Goods by Sea – Limitations on Party Autonomy

7. Louise Thorn Bøttkjær
   Votes for sale. Essays on clientelism in new democracies.

8. Ditte Vilstrup Holm
   The Poetics of Participation: the organizing of participation in contemporary art

9. Philip Rosenbaum
   Essays in Labor Markets – Gender, Fertility and Education

10. Mia Olsen
    Mobile Betalinger - Succesfaktorer og Adfærdsmæssige Konsekvenser

11. Adrián Luis Mérida Gutiérrez
    Entrepreneurial Careers: Determinants, Trajectories, and Outcomes

12. Frederik Regli
    Essays on Crude Oil Tanker Markets

13. Cancan Wang
    Becoming Adaptive through Social Media: Transforming Governance and Organizational Form in Collaborative E-government

14. Lena Lindbjerg Sperling
    Economic and Cultural Development: Empirical Studies of Micro-level Data

15. Xia Zhang
    Obligation, face and facework: An empirical study of the communicative act of cancellation of an obligation by Chinese, Danish and British business professionals in both L1 and ELF contexts

16. Stefan Kirkegaard Sløk-Madsen
    Entrepreneurial Judgment and Commercialization

17. Erin Leitheiser
    The Comparative Dynamics of Private Governance
    The case of the Bangladesh Ready-Made Garment Industry

18. Lone Christensen
    STRATEGIIMPLEMENTERING: STYRINGSBESTRÆBELSER, IDENTITET OG AFFEKT

19. Thomas Kjær Poulsen
    Essays on Asset Pricing with Financial Frictions

20. Maria Lundberg
    Trust and self-trust in leadership identity constructions: A qualitative exploration of narrative ecology in the discursive aftermath of heroic discourse

21. Tina Joanes
Sufficiency for sustainability
Determinants and strategies for reducing clothing consumption

22. Benjamin Johannes Flesch
Social Set Visualizer (SoSeVi): Design, Development and Evaluation of a Visual Analytics Tool for Computational Set Analysis of Big Social Data

TITLER I ATV PH.D.-SERIEN

1992

1. Niels Kornum
Servicesamkørsel – organisation, økonomi og planlægningsmetode

1995

2. Verner Worm
Nordiske virksomheder i Kina
Kulturspecifikke interaktionsrelationer ved nordiske virksomhedsetableringer i Kina

1999

3. Mogens Bjerre
Key Account Management of Complex Strategic Relationships
An Empirical Study of the Fast Moving Consumer Goods Industry

2000

4. Lotte Darsø
Innovation in the Making
Interaction Research with heterogeneous Groups of Knowledge Workers creating new Knowledge and new Leads

2001

5. Peter Hobolt Jensen
Managing Strategic Design Identities
The case of the Lego Developer Network

2002

6. Peter Lohmann
The Deleuzian Other of Organizational Change – Moving Perspectives of the Human

7. Anne Marie Jess Hansen
To lead from a distance: The dynamic interplay between strategy and strategizing – A case study of the strategic management process

2003

8. Lotte Henriksen
Videndeling
– om organisatoriske og ledelsesmæssige udfordringer ved videndeling i praksis

9. Niels Christian Nickelsen
Arrangements of Knowing: Coordinating Procedures Tools and Bodies in Industrial Production – a case study of the collective making of new products

2005

10. Carsten Ørts Hansen
Konstruktion af ledelsesteknologier og effektivitet

TITLER I DBA PH.D.-SERIEN

2007

1. Peter Kastrup-Misir
Endeavoring to Understand Market Orientation – and the concomitant co-mutation of the researched, the re searcher, the research itself and the truth

2009

1. Torkild Leo Thellefsen
Fundamental Signs and Significance effects
A Semeiotic outline of Fundamental Signs, Significance-effects, Knowledge Profiling and their use in Knowledge Organization and Branding

2. Daniel Ronzani
When Bits Learn to Walk Don't Make Them Trip. Technological Innovation and the Role of Regulation by Law in Information Systems Research: the Case of Radio Frequency Identification (RFID)

2010

1. Alexander Carnera
Magten over livet og livet som magt
Studier i den biopolitiske ambivalens