FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Modeling and Understanding Social Influence in Groups and Networks

DISSERTATION

zur Erlangung des akademischen Grades

Doktorin der Technischen Wissenschaften

eingereicht von

Mag.rer.nat. Julia Neidhardt
Matrikelnummer 9702295

an der Fakultät für Informatik
der Technischen Universität Wien
Betreuung: Univ.Prof. Dipl.-Ing. Dr.techn. Hannes Werthner

Diese Dissertation haben begutachtet:

Univ.Prof.i.R. Dr. Wilfried Grossmann	Prof. Dr. Markus Zanker

Wien, 29. Februar 2016

Julia Neidhardt

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Modeling and Understanding Social Influence in Groups and Networks

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktorin der Technischen Wissenschaften

by

Mag.rer.nat. Julia Neidhardt
Registration Number 9702295

to the Faculty of Informatics

at the Vienna University of Technology

Advisor: Univ.Prof. Dipl.-Ing. Dr.techn. Hannes Werthner

The dissertation has been reviewed by:

Univ.Prof.i.R. Dr. Wilfried Grossmann	Prof. Dr. Markus Zanker

Vienna, 29th February, 2016

Julia Neidhardt

Erklärung zur Verfassung der Arbeit

Mag.rer.nat. Julia Neidhardt
Favoritenstraße 9-11/188-4; 1040 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 29. Februar 2016

Julia Neidhardt

Danksagung

Ich möchte an dieser Stelle meinem Betreuer Prof. Hannes Werthner für die große Unterstützung in den letzten Jahren danken, im Speziellen dafür, dass er mich in meiner wissenschaftlichen Tätigkeit immer gefördert hat, und für die Möglichkeit, meine Zeit mit inhaltlich spannenden und herausfordernden Fragestellungen zu verbringen. Den KollegInnen in der E-Commerce Arbeitsgruppe danke ich für die gute Zusammenarbeit und die kollegiale Atmosphäre. Im Besonderen möchte ich hier meine Freundinnen Mamen Calatrava Moreno, Natalia Rümmele und Keaw Krathu hervorheben, die mich durch das Doktoratsstudium begleitet haben. Mamen danke ich auch für das Lektorat eines Teils dieser Arbeit und die aufmunternden Worte in den letzten Wochen. Natascha Zachs möchte ich für den Zuspruch und die vielfältige Unterstützung in allen Belangen danken. Am Institut für Softwaretechnik und interaktive Systeme arbeite ich gerne. Dieses Institut ist sehr international, mit KollegInnen aus vielen verschiedenen Ländern von verschiedenen Kontinenten, was ich als sehr bereichernd empfinde, vor allem angesichts zunehmender gegenteiliger Tendenzen in der Gesellschaft.

Ein wichtiger Teil meiner wissenschaftlichen Entwicklung war der mehrmonatige Aufenthalt in der SONIC Arbeitsgruppe an der Northwestern University, der mich sehr bereichert hat. Hier danke ich besonders Prof. Noshir Contractor und Yun Huang, von denen ich sehr viel gelernt habe. Aus diesem Aufenthalt ist eine fruchtbare Zusammenarbeit entstanden, die noch immer andauert. Sue und Tim Schell danke ich, dass sie mich immer bei sich aufnehmen, wenn ich in Chicago bin.

Meinen Eltern Frieda Neidhardt und Wolfgang Sagerschnig danke ich für ihre große Unterstützung in all der Zeit und dafür, dass sie es mir immer ermöglicht haben, mich meinen Interessen entsprechend zu betätigen. Vielen Dank auch meinem Freund Ernst Lammer, ohne den ich das alles nie geschafft hätte. Meinen Freundinnen Sophie Lampl, Elisabeth Günther und Saskja Schindler danke ich für viele Gespräche und die Hilfe bei soziologischen Fragestellungen.

Außerdem möchte ich Prof. Wilfried Grossmann und Prof. Markus Zanker danken, dass sie sich bereit, die Dissertation zu begutachten.

Kurzfassung

Von sozialem Einfluss spricht man, wenn eine Person ihr Verhalten basierend auf dem Verhalten von anderen Leuten im sozialem System ändert. Prozesse, die das bewirken, sind sehr komplex und schwer zu fassen. Typischerweise musste großer Aufwand betrieben werden um passende Daten für Forschungsprojekte zu finden. Heute ergibt sich jedoch ein anderes Bild. Durch neue Technologien im Allgemeinen und dem World Wide Web im Besonderen, stehen großen Mengen detaillierter Daten über menschliches Verhalten und soziale Interaktionen zur Verfügung.

Ziel dieser Arbeit ist es, Prozesse im Zusammenhang mit sozialem Einfluss im großen Rahmen zu modellieren und zu berechnen. Zunächst werden verschiedene Ansätze von unterschiedlichen Disziplinen in einem konzeptionellen Bezugsrahmen zusammengeführt. Dieser Bezugsrahmen unterscheidet drei Ebenen der Information, nämlich die individuelle Ebene, die Gruppen-Ebene sowie die Netzwerk-Ebene. Ein Ziel dieser Arbeit ist es, diese verschieden Ebenen in die Berechnungsmodelle zu integrieren.

Um jede Ebene im Detail darzustellen sowie die Unterschiede zwischen den Ebenen zu veranschaulichen, werden relevante Methoden diskutiert und empirische Studien durchgeführt. Es wird insbesondere gezeigt, wie die Gruppen-Ebene bzw. die Netzwerk-Ebene die individuelle Ebene erweitern. Das Anwendungsgebiet auf der Gruppen-Ebene stellen Reise-Recommendersysteme dar. Mit Hilfe der Geometrischen Datenanalyse wird gezeigt, wie kollektive Vorlieben genutzt werden können, um BenutzerInnen und deren Reiseverhalten zu beschreiben. Statistische Analysen zeigen, dass diese Darstellung das Verhalten der BenutzerInnen zutreffend beschreibt. Die Prozesse, die hier untersucht werden, basieren auf sozialem Einfluss durch Vergleich. Im Gegensatz dazu ist der Fokus auf der Netzwerk-Ebene sozialer Einfluss durch Kommunikation. Zuerst wird Churn-Verhalten, also das Abwandern von KundInnen, in einem Multiplayer Online Spiel analysiert, wobei sozialer Einfluss zwischen den SpielerInnen berücksichtigt wird. Um Nachteile herkömmlicher Modelle zu umgehen, werden Modelle basierend auf Conditional Random Fields eingeführt. Danach wird untersucht, ob die Stimmungen von BenutzerInnen in einem Online-Reiseforum miteinander verbunden sind. Bei diesen Modellen liegt der Fokus darauf, Struktur und Inhalt der Diskussion im Forum gemeinsam zu modellieren. Abschließend wird diskutiert, wie alle drei Ebenen in einem Modell integriert werden können. Dazu werden komplexe Team gegen Team Wettkämpfe in einem Multiplayer Online Spiel analysiert.

Die Resultate dieser Arbeit können zwei Kategorien zugeordnet werden: 1) Beiträge zur Verbesserung der Methodik sowie 2) konkrete Aussagen in den Anwendungsgebieten. Es wird gezeigt, dass die eingeführten Modelle sozialen Kontext zutreffend darstellen. Auch können die meisten der Modelle gut mit großen Datenmengen umgehen. Außerdem ermöglicht es das Zusammenführen verschiedener Ebenen in einem Modell, diese Ebenen und ihren Zusammenhang mit den untersuchten Einflussprozessen direkt zu vergleichen. Das führt zu umfangreicheren Einblicken in das jeweilige Anwendungsgebiet.

Abstract

Social influence occurs when a person changes her behavior according to the behavior of other people in the social system. These mechanisms are very complex and hard to capture, and traditional research on social influence had to put a lot of effort into gathering suitable data. However, today the situation has drastically changed. Due to new technologies, in particular the world wide web, vast amount of detailed data on human behavior and social interactions are available.

The main objective of this work is to capture social influence processes in computational models at a large scale. First, a framework is introduced that summarizes approaches from various fields. This framework distinguishes three levels of information (i.e., individual, group and network level) and the aim is to integrate these levels into the social influence models. To illustrate each level in detail and to show their differences, relevant methods are discussed and empirical studies are conducted. Particularly, it is shown how the group level as well as the network level extend the individual level. The application domain at the group level are travel recommender systems. With the help of geometric data analysis it is demonstrated how collective preferences can be used to describe users and their travel behavioral patterns. Statistical analyses show that this representation captures user behavior in an accurate way. Here, comparison-based social influence mechanisms are studied. At the network level the focus are communication-based social influence processes. First, churn behavior in a multiplayer online game is analyzed with the aim to control for social influence among the players. To address shortcomings of conventional models, conditional random field models are introduced. Second, interdependencies between user sentiments in an online travel forum are studied. Here, the focus are models that integrate both structure and contents of user discussions. Finally, it is discussed how all three levels can be integrated into one model. A complex setting related to team-vs-team competitions in a multiplayer online game is analyzed.

The results of this work are related to two categories: 1) methodological advances, and 2) concrete statements in different domains of application. It is shown that the introduced models are able to capture social context in an accurate way. Most of them, moreover, scale very well. Furthermore, integrating different levels of information allows to directly compare the different levels and their associations with the studied social influence processes. Thus, a more comprehensive picture of the respective domain of application is obtained.

Contents

CHAPTER 1

Introduction

1.1 Motivation

The behavior of a person is typically strongly affected by the interactions of this person with other people and her position in the social system. Both determine considerably what information a person can access and what behaviors and social norms she is exposed to. A social influence process occurs when individuals adapt their behavior, attitudes or beliefs to the behavior, attitudes or beliefs of other people in the social system. Overall, the motives of a person for adapting or for resisting to conform are manifold, but can be related either to the aim to ensure a coherent and favorable self-concept, to establish and to maintain satisfactory social connections or to response effectively to their social environment [CG04, Woo00]. Influence is not restricted to direct communication and it does not matter whether the influence process is intentional or not intentional [Lee02].

Thus, social influence mechanisms are very complex and hard to capture. Traditional research on such processes had to put a lot of effort into gathering suitable data. Typically, the focus were small scale settings including families, schools, working environments, doctors and politicians [Ras07], and the studies had to rely on self-reported data on behavior, opinions and also relationships. However, today the situation has drastically changed. Due to new technologies, in particular the World Wide Web, detailed data on human behavior and interactions is available. This data comprises both information on structure and content [LPA+09].

In the last two decades, the Web and online social communities have become increasingly important in almost all areas. The Web is not only reflecting society, it is also changing it at different levels. Not only private communication is more and more taking place online but also business interactions are increasingly performed on the Web. It enables and mediates new forms of interactions in an unprecedented way. For a variety of activities, for instance, physical proximity is less important today, e.g., teaching and knowledge transfer

is increasingly happening via Web platforms; and it has been shown that both in academic and in business settings, teamwork can be done effectively online [FBR99, MC00].

Together with the success and the tremendous growth of the Web, also the amount of static and dynamic data that is permanently generated has exploded. Along with all the risks and challenges in this context that society has to face (e.g., ownership of the data and privacy issues), great opportunities to study individual and group behavior arise. New disciplines emerge that are inherently interdisciplinary. Computational social science, for instance, makes use of massive amounts of data that captures fine-grained interaction and communication patterns as well as locations of millions of people. Also longitudinal data is collected and analyzed at a large scale. All this can lead to qualitatively new insights into human interactions and collective behavior. However, also methodological challenges come along with this development. Traditional methods to model and analyze human behavior and social networks have been developed for much smaller and static data samples. Thus, new methods and methodologies are required [LPA+09]. Furthermore, the term Web Science has been established as expression for an interdisciplinary research field that focuses on both the micro level (e.g., small technological innovations) and the macro level (e.g., large scale phenomena that affect society and commerce globally) of the Web [HSH+08, THC+15].

Also social influence processes are increasingly studied and discussed in the context of the Web and Web related applications. This leads to new perspectives on the mechanisms of spreading phenomena. The diffusion of information and behaviors can now be examined on a large scale. Qualitatively new insights into efficient and effective communication patterns can be gained taking into account both structure and content. Also how people react to external influence mechanisms based on their intrinsic motivations can be compared in novel ways. Thus, a number of disciplines including sociology, behavioral sciences, psychology and marketing can benefit from these now possibilities to a great extend. Research results, moreover, are directly applicable to societal, political and commercial activities.

Governmental institutions and policy makers with the goal to promote a certain behavior, for example, healthy or eco-friendly behavior, are interested in how to target the right people and how to address them appropriately. Recently, the term *nudging* has been established in this context. It is based on the idea that small changes in design can impact individual behavior in a "positive" way so that there is a benefit for both the individual and the society [ST09]. Although this form of *libertarian paternalism* is clearly perceived controversial, the approach has been applied by governments and policy makers [HW10]. Recently the idea is more and more extended to *big nudging*, i.e., the development of influencing mechanisms based on large-scale data analysis [HFG+16]. Of course also companies want to know how to target their costumers in the right way in order to distribute their products and services and to increase their revenue. There is an increasing interest of the bussiness sector to collect and analyze relevant data. Data mining techniques and predictive modeling move more and more into the focus of customer segmentation strategies and customer relationship management [TC11, Cho15].

Insights can also be utilized in the context of recommender systems, which not only recommend but also aim to persuade people [GF06].

Summing up, new technologies and new sources of data provide unprecedented opportunities to analyze human behavior and interactions and to gain qualitatively new insights into social influence processes. To conduct these large-scale analyses, new models and advanced techniques are required. Thus, to develop such approaches, disciplines including computer science and statistics are demanded. At the same time, this research is inherently interdisciplinary as theories from the social sciences and the humanities are needed to model and to understand collective human behavior and social context. This interdisciplinarity is a challenge on its own, as different fields have different research cultures as well as terminologies.

However, as a researcher, one has to be aware that data combined with the right technology is a powerful instrument to exert influence on people and also to manipulate them. Research in this area can become very delicate and ethical guidelines and transparency might be particularly crucial. This has also been shown by the controversy about a recent Facebook study on emotional contagion. The study tried to influence the emotions of the Facebook users by filtering their news feeds in certain ways without asking the participants to give their consent. One further aspect that was strongly criticized afterwards, was the fact that emotions, i.e., something particularly personal and intimate, were manipulated [CPD15].

1.2 Problem Statement

Individuals form their opinions, choose their behaviors, and imitate others in a complex environment in which typically disagreeing opinions and different behaviors exist. Thus, social influence processes are permanently present. However, they are hard to understand and extremely difficult to operationalize. It is very challenging to develop mathematical, statistical and computational models of individual behavior that at the same time take social context into consideration; such models are highly complex and typically do not scale. There is no integrated theory of social influence but various approaches in the social sciences, psychology and also marketing exist. We give an overview of some of these concepts and typologies in Chapter 2. This overview, which is by no means exhaustive, should demonstrate that our work draws on a number of disciplines, which makes the topic inherently interdisciplinary.

The overall goal of our work is to understand social influence processes. Today, due to all the data that are available, the models that exist and the potential application scenarios, questions addressing such mechanisms are obviously both clearly relevant and fascinating to study. In our work, we aim to capture social influence processes within abstract models. In order to operationalize these mechanisms, we focus in our analysis on the following aspects:

- Based on [Lee02] we categorize social influence processes as *communication* and *comparison processes*.
- Furthermore, as we are interested in human interactions mediated by Information and Communications Technology (ICT) infrastructure we relate our discussion to the *Framework of Digital Infrastructure* introduced in [WASC+15].
- We focus on behavior. As opposed to attitudes or beliefs, a behavior is typically more explicit and easier to observe [Lee02]. Thus, hereafter we will mainly refer to social influence mechanisms that affect individual behavior. However, in principle the introduced approaches are in the same way capable of modeling changes in attitudes or beliefs.

1.2.1 Social Influence: Communication and Comparison

Social influence occurs when a person changes her behavior according to the behavior of other people in the social system. In general, the influence process can be intentional or non-intentional. Furthermore, social influence mechanisms are not restricted to direct communication, but information about relevant others has to be available [Lee02, RPE01]. Overall, social influence processes can be subsumed as *communication* and *comparison* processes [Lee02]:

- *Communication:* Friends and acquaintances are typically crucial for a person to develop preferences and to form opinions. Social relations, moreover, have an important impact on the behavior of a person. These social influence processes are enabled through direct communication of an individual with her so-called *social frame of reference*.
- *Comparison:* An individual, when building up her social identity, is not only influenced by her immediate social surroundings but more generally by society as a whole. The individual compares herself to relevant others, i.e., persons who are perceived as similar, e.g., who have the same gender, age, or occupation. These relevant others form her social frame of reference; the behavior exhibited by this social frame of reference is considered as appropriate for somebody like her or somebody in her position. Here, social influence processes take effect through comparison. This is also related to the concept of *habitus* of Bourdieu [Bou84]. Broadly, habitus summarizes the way a person acts, thinks, dresses, speaks and gesticulates and also comprises a person's taste. Importantly, this habitus results from both individual disposition and social structure, two dimensions that mutually reinforce one another. Individuals internalize the social structure, which then is the basis for how individuals perceive themselves and the social space. It determines which behaviors are correct and which are out of the question.

Typically, if social influence occurs, both communication and comparison processes are present, and they are hard to distinguish empirically.

In our analysis we focus on human interactions enabled through ICT infrastructure in general and the Web in particular. Thus, we consider the *Framework of Digital Infrastructure* with its five layers: (1) *individual*, (2) *group/social*, (3) *corporate/enterprise*, (4) *network/industry*, and (5) *government/policy* [WASC+15].

The relevant layers for our discussion are the *individual layer*, where the focus is on a user and her interaction with ICT devices an services, and the *group/social layer*, where the focus is on how technology enables group interactions and offers opportunities for groups. Here, a group is defined as a collection of individuals who may or may not interact. Thus, the group layer comprises both individuals, who interact, and individuals, who share certain characteristics but do not necessarily interact. However, in our discussion on social influence we distinguish these two aspects. Thus, we obtain three *levels of information*:

- The *individual level* capturing information on independent individuals;
- The *group level* capturing information on sets of individuals;
- The *network level* capturing information on social network structure.

We do not consider the network/industry layer within the *Framework of Digital Infrastructure* since here market structure rather than social network structure is addressed. Furthermore, we do not discuss the layers corporate/enterprise and government/policy in detail. However, actors belonging to those layers could apply the results of our analysis as they might have interests in exerting social influence and pursuing others (e.g., related to our discussion in Section 1.1).

Summing up, we follow in our work the approach introduced in [Lee02] and differentiate social influence mechanisms into communication and comparison processes. However, whereas in [Lee02] the focus is exclusively on the *network level*, we extend the discussion and also consider the *individual level* as well as the *group level*.

Now we discuss the three levels of information in more detail and with respect to different perspectives. In Table 1.1 an overview is presented that integrates different aspects. It can be considered as an *ontological framework of social influence* as it summarizes *"what exists that we might acquire knowledge of"* ([Hay02], p. 61) and also comprises methodological strategies.

1.2.2 Individual Level

With respect to the *Framework of Digital Infrastructure*, this level focuses on individuals and their interactions with ICT devices and services. Every person is considered as independent, social context is not explicitly taken into consideration. Here, individual characteristics are of interest. To understand the individual motives of a person to adopt a behavior, psychological theories apply. In general, such motives can be subsumed under

	Framework of Digital Infrastructure	Social Sciences; Theoretical Background	Empirical Approach	Mathematical Abstraction; Measures
Individual Level	Individual Layer	Psychology; Social Influence: Individual Motives	Regression Models	Vectors; Attributes
Group Level	Group Layer: Focus Similarity	Sociology; Social Influence: Communication and Comparison	Geometric Data Analysis (GDA)	Euclidean Space; Associations, Affiliations and Comparison
Network Level	Group Layer: Focus Interactions	Sociology; Social Influence: Communication and Comparison	Social Network Analysis (SNA)	Networks; Relations

Table 1.1: *Social Influence Framework*: ontological framework that relates different aspects of social influence to different levels of information.

the aspects accuracy, affiliation and self-concept and are related to intrinsic motivations of a person.

Mathematical models that are applied at this level also regard the individuals as independent. To capture the characteristics of a person, attributes are used. In our discussion these attributes do not only refer to non-changeable individual characteristics such as gender but also to mutable variable such as individual preferences or behaviors; following the approach of [Lee02]. To get insights into the overall behavior in the social system, the distribution of the attribute might be considered, e.g., number of smokers and non-smokers. Typically associations between attributes are of interest in order to find out which individual characteristic foster a certain behavior; e.g., whether there is an association between gender and smoking. The correlation coefficient can be used to quantify the relation and statistical inference helps to asses the association. Thus, to study the behavior, we can apply methods such as descriptive statistics, t-tests, and, in particular, regression models.

Thus, although we might examine a high number of individuals, no interdependencies between them are considered. The focus of the analysis are dependencies of one individual attribute, representing for example a behavior, on one or more other attributes of the same individual. Social context is not explicitly modeled.

1.2.3 Group Level

At this level of information, groups of individuals are considered and thus social context is addressed. However, we do not take into account whether the individuals interact, we rather focus on similarities between them. Social influence is mainly discussed with respect to comparison mechanisms taking into account sociological theories. Of course,

also communication based social influence processes might occur but in general they cannot be identified as interactions between the individuals are not tracked at this level.

To capture social context, Geometric Data Analysis (GDA) is used [LRR04]. Knowing, which social groups a person belongs to, allows implications about her behavior. These groups are determined by associations among categories or individual attributes, e.g., affiliation of the individuals or their preferences. GDA aims to detect latent dimensions that capture collective behavior and preferences by examining these associations. The identified dimensions characterize the setting as they form the basis of a metric space, in which the individuals are located. The positions in the space have a meaning, and individuals or groups that are close are more similar than individuals or groups that are further away. A goal of GDA is to identify groups of users with shared characteristics. Often cluster analysis techniques are applied to dedect such *clouds of points* [LRR04].

Methods that are used to determine the dimensions and to construct the metric space include correspondence analysis, factor analysis and principal component analysis.

GDA is clearly relevant for our discussion on social influence as it establishes connections between the characteristics of a person and her position in the social space. The assumption is that this position strongly impacts individual behavior as it determines, which social norms one is exposed to. A person compares herself with others who are close in the metric space. To model and analyze the social space in such a way has a noteworthy tradition in sociology. In particular Pierre Bourdieu, who strongly relied on correspondence analysis in his empirical social studies, established this approach [BS13].

1.2.4 Network Level

At the network level interactions are considered. Thus, individuals are explicitly regarded as interdependent. Here, both communication and comparison mechanism are observable.

To address social context, Social Network Analysis (SNA) is applied [WF94]. This approach considers relations such as friendship, co-working connections or communication ties as crucial to understand the social system in general and the individual and her preferences in particular. Relations between the individuals enable influence processes. Furthermore, the position of an individual in the network structure strongly determines her opportunities as well as limitations.

Social networks are typically formalized as graphs in which the nodes represent individuals and the edges relations between them. Additional information can be assigned to both nodes (e.g., individual attributes) and edges (e.g., weights to quantify the strength of a tie). In terms of social network analysis, social influence expresses that given the edges (i.e., connections) between the nodes (i.e., individuals), the nodal attributes (i.e., behavior of the individuals) are influencing one another (i.e., the behavior is contagious). As we will see, it is quite challenging to verify whether social influence mechanisms in fact occur in a network.

There are two mechanisms how network structure can foster social influence: First, a person might be influenced by her immediate network connections. The behaviors of these peers impact the behavior of the individual due to direct interactions or communication. Second, a person might be influenced by other individuals in the network that are in the same structural position. They might not directly communicate but they might compare and imitate their behaviors and learn from one another as they occupy the same social role facing the same challenges.

Figure 1.1 illustrates the differences between the three levels of information with respect to communication and comparison processes. The perspective of one individual (i.e., *ego*) is related to her social frame of reference (i.e., *alters*). Only the network perspective allows for a structural distinction between the mechanisms.

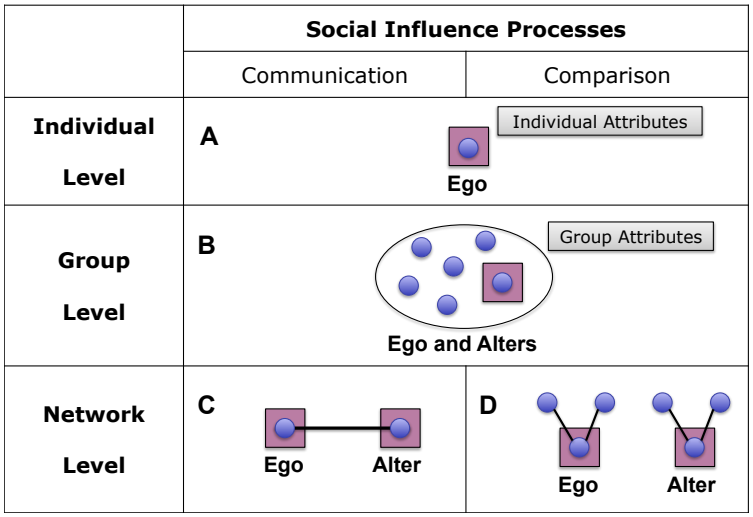

Figure 1.1: Social influence processes and levels of information. The network perspective allows for a structural distinction between communication and comparison mechanisms.

1.3 Research Questions

In Section 1.2 we discuss different perspectives on social influence mechanisms. In particular, we distinguish three levels of information, each of which with a different focus and different approaches. To gain a comprehensive view on social influence we are interested in combining insights from different levels and different fields. Therefore, we address the following research questions in this thesis:

- **Research Question 1 (RQ1)**: Are there computational frameworks that integrate different levels and approaches when studying social influence on a large scale?
- **Research Question 2 (RQ2)**: What do we gain by taking different levels of information into account when studying social influence phenomena?

1.3.1 Focus of This Work

The overall focus of this work is to model and to understand social influence process in order to answer **RQ1** and **RQ2**.

First, literature from different scientific fields is studied to get insight into distinct ontological and methodological approaches to this topic. These approaches are integrated into a *Social Influence Framework* (see Table 1.1) that comprises three levels of information.

To illustrate each level in detail and to show their differences, both a methodological overview is given and empirical examples are used. In particular, we examine how the group level on the one hand and the network level on the other hand extend the individual level. Finally we discuss how to integrate all the three levels of information. As it is very challenging to operationalize social influence, we make the following simplification: In our models, we only capture whether or not a person is affected by social influence but we do not take her individual motives into consideration, i.e., we do not distinguish whether the motives are related to accuracy, affiliation or self-concept. Furthermore, as mentioned in Section 1.2, we focus on modeling behavior rather than attitudes or opinions as behavior is easier to observe.

To exemplify the main concepts data from literature on teenage smoking behavior is used [MW96]. This data set is appropriate and serves our purposes as social influence processes among the teenagers have already been identified [SSP10]. The introduced concepts are applied in empirical studies considering the group level, the network level as well as all levels together.

The application domain at the group level are travel related recommender systems. Here comparison processes are studied. We use the GDA approach to propose user-centric models rather than focusing on product features. Statistical analyses are conducted to show that this representation captures user behavior in an accurate way. At the network level our focus are communication based social influence processes. First we analyze churn behavior in a multiplayer online game with the aim to control for social influence among the players. As existing models do not scale, we propose an approach based on conditional random fields, i.e., undirected graphical models. Furthermore we study interdependencies between the sentiments of users on a travel online forum. Here, we focus on a model to integrate both structure and contents of user discussions. Finally, we study the impact of different levels of information on team performance in a complex head-to-head setting. Here, the challenge is to model two outcome dimensions, i.e., winning the match and duration of the match. Therefore, new measures are introduced. As it is an empirical question, whether or not social influence occurs, we distinguished two layers when the different research steps: The first layer comprises mathematical, statistical and computational models to capture social influence mechanisms. The second layer is about the application of the models to empirical questions in order to obtain concrete statements about different domains. This was also how we organized our analysis.

1.4 Methodological Approach

Regarding the methodological approach, there are two layers that can be distinguished:

Focus of the first layer are *mathematical, statistical and computational models* to capture social influence mechanisms. Here, advantages and limitations of such models are discussed and compared with respect to the introduced *Social Influence Framework* (see Table 1.1) that allows to differentiate various perspectives and different levels of information. Also the challenges at each level are identified. As discussed in Section 1.2, the overall objective is to investigate how the behavior of an individual is influenced by her social frame of reference. This implies the following: As opposed to standard classification or regression tasks where only information on individual attributes are considered, we explicitly need to build social context and interdependencies into our models. On the group level this is done by using the behavior and preferences of all individuals in order to construct an abstract metric space. On the network level, we use statistical network models to explicitly account for interdependencies between the individuals. To address some of the shortcomings of traditional approaches, we introduce conditional random field models as a novel way to capture social influence in networks. Furthermore, we discuss how to account for different levels of information within one model. Here, new measures are introduced. Furthermore applying a comprehensive methodology, we also aim to take theories from the social and behavioral sciences into consideration when developing our models and measures. The methods and theories that we build upon are described in Chapter 2.

The second layer is about *empirical analyses* using the introduced models and approaches. In several case studies concrete models are developed and iteratively refined to find out about the strengths of the different approaches as well as their shortcomings. To quantify, moreover, the power of social influence models in several domains, the results based on the introduced approaches are compared to results obtained by conventional approaches, e.g., standard logistic regression models. Thus, based on our work one could also answer concrete sociological questions, but this is not the focus. What we want is to develop methods that can be used to solve *real world* problems.

As it is an empirical question, whether or not social influence occurs, both layers strongly interact. Advances on the theoretical layer have to be evaluated empirically.

Furthermore, one can refer to the *Design Science Research* (DSR) *Framework* introduced in [HMPR04]. Using this elaborated framework, we meet its seven guidelines (see Table 1.2):

- Our *artifacts* are methods and models as well as their integration into a coherent framework.
- *Problem relevance* is given by the omnipresence of social influence phenomena and the need to understand them. High volumes of dynamic and very detailed data that reflect human interactions are permanently generated by new data sources

including the Web and online social communities. Models and methods to analyze this data can provide valuable insights into human behavior and social interaction as well as related phenomena in this context.

- *Evaluation* of the models and methods using quantitative and qualitative techniques: the models are assessed quantitatively by appropriate measures (goodness-of-fit, prediction rate) using suitable data sets; the derived statements about the different domains are compared to knowledge gained through the study of literature.
- *Contributions of our work* are methods and models that extend the knowledge base (i.e., the theoretical foundations and research methodologies). Furthermore, we apply existing knowledge in new contexts and novel ways. Our main contributions are described in Section 1.5.
- The *rigor of research* is given by its grounding in well developed theories of mathematics, statistics and the social sciences. Our models are developed in an iterative way (identification of appropriate sub samples, selection of variables, number of factors and clusters, development of new measures) with respect to both the accuracy of the models and their interpretability.
- In this *search process* results from literature, experience, creativity and intuition are used to achieve improvements.
- When *communicating our research* we not only address an audience that is familiar with mathematical and statistical methods but also aim to show the usefulness and applicability of our research to problems in an interdisciplinary context.

1.5 Main Results

Whether or not a certain behavior is in fact influenced by social context is an empirical question. Thus, different approaches are tested, adapted and combined using case studies. As a consequence, our results are related to two categories: (1) Methodological advances, and (2) concrete statements in different domains of application.

1.5.1 Methodological Advances

We now summarize the methodological advances our work contributes to. These results are related to **RQ1**.

Ontological Framework

M1: Social Influence Framework. There exist various approaches to describe and categorize social influence mechanisms in literature (see Chapter 2). To approach this topic systematically and to operationalize social influence we introduce a framework to capture various ontological and methodological aspects. This is a challenge on its

Guideline	Description
Guideline 1: Design as an Artifact	*"Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation".*
Guideline 2: Problem Relevance	*"The objective of design-science research is to develop technology-based solutions to important and relevant business problems".*
Guideline 3: Design Evaluation	*"The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods".*
Guideline 4: Research Contributions	*"Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies".*
Guideline 5: Research Rigor	*"Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact".*
Guideline 6: Design as a Search Process	*"The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment".*
Guideline 7: Communication of Research	*"Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences".*

Table 1.2: DSR Guidelines ([HMPR04], p. 83).

own due to the breadth and depth of this topic in different scientific fields with their own terminologies and ontologies. Our *Social Influence Framework* distinguishes three levels of information (see Table 1.1). To illustrate the differences between the three levels, we use data from literature to compare the three levels explicitly. Furthermore, we demonstrate how to integrate different levels within one model.

Integration of Methods within the Framework

M2: Abstract Metric Space for Recommendations. Based on this framework we introduce a new perspective on user modeling in the context of personality-based recommender systems. Moving away from product-features and rating behavior at the center of such models, we focus on users, their preferences and their social context. Here, preferences are determined by both travel behavioral patterns and personality traits. By using collective preferences we construct abstract metric spaces to model behavior and to deliver recommendations. To elicit the preferences of a user, we introduce an innovative picture-based approach. In this work we present a thorough statistical analysis demonstrating that these user models are both meaningful and capable of representing the setting in an accurate way. Several distinct groups are detected in which users exhibit normative behavior. Thus, these groups can be targeted by a recommender system.

M3: Conditional Random Field (CRF) Models. When dealing with social influence processes in networks, models are needed that are able to capture interdependencies between the users. There exist only very few cross-sectional models in the literature that meet this requirement. These models, while being well applicable to small networks, do not scale. As we are interested in analyzing larger settings we introduce Conditional Random Field (CRF) Models as a novel way to capture social influence in networks. Compared to traditional network models for social influence they scale very well and can capture complex interdependencies between the individuals.

M4: Network Models Combining Structure and Content. Utilizing detailed data on user discussions in an online forum, we develop network models that capture interdependencies between the sentiments of the users. These sentiments are obtained by applying text mining techniques and sentiment analysis to the user comments. The network structure reflects communication ties in the forum. Thus, by combining structure and content, we show that social influence models can be used to study emotional contagion.

M5: Relative Performance Models. To study both outcome and duration of head-to-head competitions of two teams, we develop models of relative performance. To capture the complex setting, factors related to distinct levels of information are constructed. We introduce different ways to assess the relative importance of each of the factors with respect to the winning behavior of a team and the duration of a match.

1.5.2 Application Domains

By applying the introduced models in empirical studies, we obtain concrete insights into social influence processes. These results are related to **RQ2**.

Empirical Statements

A1: Travel Recommender Systems. We study the positions of the users in the space spanned by the seven identified travel related factors *Sun & Chill-Out*, *Knowledge & Travel*, *Independence & History*, *Culture & Indulgence*, *Social & Sport*, *Action & Fun* and *Nature & Recreation*. Statistical analyses are conducted that show that there are significant differences between the age groups of the users with respect to the seven factors. The same applies for gender. We conduct a cluster analysis and detect six groups of users and show that also the clusters are significantly different with respect to the factors. Thus, users belonging to a certain group exhibit a distinct travel behavior than users belonging to another group.

A2: Churn Analysis. We investigate churn behavior in the Massively Multiplayer Online Role-Playing Game (MMORPG) EverQuest II [Son16] in order to find out, which factors influence quitting behavior. Based on literature, hypotheses are phrased. The results imply that the commitment of a player, her achievements and community effects

decrease the likelihood that this player will quit the game. However, those results are not consistent across the different models. On the other hand, all models detect a significant contagion effect, i.e., if a partner leaves the game a player is also more likely to quit the game as well.

A3: Sentiment Contagion in Travel Online Forums. We study a travel online forum, where users meet and interact before attending a group tour together, with the goal to find out whether the emotions of the travelers are interrelated. The emotions of a user are determined based on her comments. The results imply that the emotions are in fact interrelated. Furthermore, also individual attributes such as gender have an impact on a user's emotions.

A4: Team-vs-Team Competitions. We analyze duration and outcome of team-vs-team competitions in the Multiplayer Online Battle Arena (MOBA) game Dota 2 [Val16]. Here, the teams' overall goal is to beat the opponent. We test hypotheses related to the skills of the players, to past co-relations within the team and to experiences with outside players. Our results show that teams with high skills players with more previous collaboration are more likely to win and win faster. On the other hand, teams with members who played in many teams before are more likely to win but it will take longer time.

1.5.3 DSR Knowledge Contribution Framework

In Figure 1.2 we show how these results can be related to the *DSR Knowledge Contribution Framework* [GH13]. This framework allows to classify research contributions with respect to four quadrants that capture the maturity of the problem, i.e., the application domain, on the one hand and the maturity of the solution, i.e., the knowledge or artifact, on the other hand (see Figure 1.2). Based on this, the four quadrants represent *improvement*, *invention*, *routine design* and *exaptation* ([GH13], p. 345):

- *Improvement:* in this quadrant new solutions for known problems are developed.
- *Invention:* in this quadrant new solutions for new problems are invented.
- *Routine design:* in this quadrant known solutions are applied to known problems.
- *Exaptation:* in this quadrant known solutions are extended to new problems (e.g., adopting solutions from other fields).

Thus, based on this definition, we classify our results related to the methodological advances **M4** and **M5** as well as to the empirical results **A2**, **A3** and **A4** as *improvements* as they all constitute new solutions for known problems (either related to improvement of the method or related to the improvement of the empirical approach).

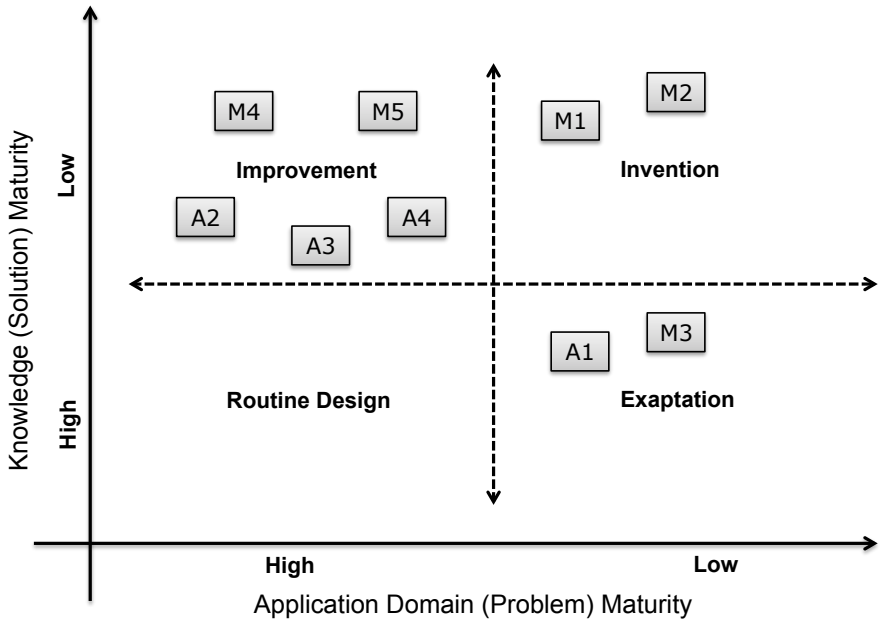

Figure 1.2: Classification of results within the *DSR Knowledge Contribution Framework* [GH13].

We relate the introduced *Social Influence Framework*, i.e., result **M1**, and the picture-based approach to recommeder systems, i.e., result **M2**, to the category *invention* as both constitute new solutions for new problems. The *Social Influence Framework* aims to combine knowledge of various fields in a systematic way to facilitate the development of new solutions while at the same time specifying the problem. Therefore, it can be considered as *"an exploratory search over a complex problem space"* ([GH13], p. 345). Personality-based recommender systems are new problems that we address applying a new approach, i.e., the picture-based approach to preference elicitation.

Finally, the adoption of Conditional Random Field Models, i.e., result **M3**, to studies of social influence can be seen as *exaptation*, as these models are known solutions for certain large-scale data mining problems but have to our best knowledge not been applied in this way to social influence problems before. We also relate the personality-based travel recommender systems. i.e., result **A1**, as *exaptation*. Personality-based recommender systems are a new problem that we address applying a known solution, i.e., a GDA based approach.

1.6 Structure of This Work

This work is structured as follows: In Chapter 2 we present an overview of theories of social influence and other relevant literature. Furthermore, essential concepts and reviews are introduced.

In Chapter 3 we study social influence processes on the group level. A main goal is to analyze how this level extends the individual level. To illustrate the differences we use a data set from literature on teenage smoking behavior. At the individual level this smoking behavior is modeled with the help of logistic regression. Different variables are compared regarding their predictive power. In the next step, the setting is modeled applying the GDA approach. Factor analysis is used to construct a lifestyle space and clusters of teenagers in this space are determined and analyzed. Furthermore, the differences to the individual level are discussed. Then, we apply the same group level approach in the context of picture-based travel recommender systems. First we introduce the approach in detail. After this, we conduct thorough statistical analyses to show that the representation of the users with respect to seven basic factors capturing distinct travel behavioral patters is both reasonable and accurate. This is also confirmed by a cluster analysis where distinct groups of users exhibiting different travel behaviors are identified. Thus, comparison mechanisms are likely to shape the social space.

In Chapter 4 we study communication based social influence processes at the network level. Existing cross-sectional models to capture social influence in networks are compared again in the context of the teenage smoking data set. To address some of the limitations of these models, we introduce Conditional Random Field Models as a novel way to study social influence phenomena in networks. Furthermore, we discuss how the analysis on the network level extends the individual level. Then, we apply the same approach to model and predict churn behavior in an online community. In particular, we want to find out, which factors retain users and which factors make them leave. Next, we analyze structure and content of user discussions in an online travel forum with the goal to find out whether the sentiments of the users are interrelated. Further individual attributes that impact the sentiment of a user are identified. Finally, we discuss social influence and social selection with respect to the introduced models.

In Chapter 5 multi-level analyses are conducted. Utilizing the example of smoking teenagers again, we illustrate how information at the group level can be included in network models. Then we focus on complex team-vs-team settings, where we consider two outcome variables, i.e., winning or losing, on the one hand, and the duration of a match, on the other hand. To predict the winning behavior of a team factors related to distinct levels of information are constructed, i.e., related to the composition of a team capturing the individual level, related to the relations within a team capturing the network level and related to the position of a team within the ecosystem of teams capturing the group level.

Finally, in Chapter 6 conclusions are presented, open issues are discussed and future work is outlined.

This work is based on the following peer-reviewed papers:

- Julia Neidhardt, Rainer Schuster, Leonhard Seyfang, and Hannes Werthner. Eliciting the users' unknown preferences. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 309–312, 2645767, 2014. ACM

- Julia Neidhardt, Leonhard Seyfang, Rainer Schuster, and Hannes Werthner. A picture-based approach to recommender systems. *Information Technology and Tourism*, 15(1):49–69, 2014
- Julia Neidhardt, Nataliia Rümmele, and Hannes Werthner. Can we predict your sentiments by listening to your peers? In *Information and Communication Technologies in Tourism 2016*, pages 593–603. Springer International Publishing, 2016
- Julia Neidhardt, Yun Huang, and Noshir Contractor. Team vs. team: Success factors in a multiplayer online battle arena game. In *Academy of Management Proceedings*, volume 2015, 2015

Furthermore, some of the work presented in this thesis has not been published yet but is in the process of being finalized and getting submitted. However, it has been presented to the relevant community in the following talks (presenter: Julia Neidhardt):

- Julia Neidhardt, Yun Huang, Hannes Werthner, and Noshir Contractor. Conditional random field models as a way to capture peer influence in social networks. `http://www.ec.tuwien.ac.at/neidhardt/Sunbelt2015.pdf`, June 2015. Sunbelt XXXV, Brighton, UK
- Yun Huang and Julia Neidhardt. From networks to space: Constructing metric spaces for social interactions. `http://www.ec.tuwien.ac.at/neidhardt/EISSII.pdf`, October 2015. Empirical Investigation of Social Space II, Bonn, Germany

Picture-based recommender systems are a major research focus of the E-commerce group at TU Wien; the study on recommender systems presented in Chapter 3 is part of this bigger project. Furthermore, the work on conditional random field models (see Chapter 4) and on team-vs-team models (see Chapter 5) was initiated during a research stay at the SONIC lab at Northwestern University, USA. The collaboration on these topics is ongoing. A complete list of publications can be found at the end of this work.

CHAPTER 2

State of the Art

In this chapter we present an overview of theories of social influence as well as methods that are used to study this influence. As social influence processes describe complex phenomena related to human behavior and interactions, this topic is inherently interdisciplinary. The presented overview is by no means exhaustive but we aim to demonstrate that our work draws from a number of research fields, as shown in the framework in Table 1.1 in Section 1.2. Based on this framework, we first discuss social science theories in the context of social influence, and then empirical methods and frameworks. Finally, we characterize the application domains that are relevant for the following chapters.

2.1 Theories of Social Influence

2.1.1 Individual Motives

The study of social influence in psychology is typically concerned with motives why an individual accepts social influence or resists it, which of course is not necessarily a prozess that the individual is conscious of. In the literature, different typologies of motives have been introduced. Contemporary approaches often show a differentiation into three types of motives, i.e., individuals accept social influence due to *"normative concerns for (a) ensuring the coherence and favorable evaluation of the self, and (b) ensuring satisfactory relations with others given the rewards/punishments they can provide, along with an informational concern for (c) understanding the entity or issue featured in influence appeals"* ([Woo00], p. 541).

An early framework in this context has been introduced by [Kel58]. Here, three qualitatively different mechanisms are distinguished, where all lead to the same result, i.e., a person accepts the influence and conforms:

- *Compliance* occurs when a person adopts a certain behavior either to get approval from others or to avoid disapproval. The reason for the behavioral change is due to opportunistic reasons rather than believing in this behavior. Thus, the individual gains satisfaction from the *social effects* of conforming.

- *Identification* occurs when a person adopts a certain behavior to create or maintain a relation to another individual or to a group. The behavior is strongly associated with the relationship. Accepting the influence is more crucial than the actual behavior itself. Thus, the individual gains satisfaction from the *act of conforming*.

- *Internalization* occurs when a person adopts a certain behavior because it is congruent with her existing values. Thus, the individual gains satisfaction from the *content* of the behavior.

Furthermore, the likelihood that a person accepts the influence can be considered as a combination of the three processes *"(a) the relative importance of the anticipated effect, (b) the relative power of the influencing agent, and (c) the prepotency of the induced response"* ([Kel58], p. 53).

In [CG04] individual motivations that lead people either to compliance or to conformity are discussed based on recent studies on this topic. *Compliance* refers to a behavioral change of a person as a result of a either explicit or implicit request from another person. *Conformity*, on the other hand, refers to a behavioral change in order to match the responses of others or to fit social norms. For each compliance and conformity three types of motives are distinguished, i.e., *accuracy*, *affiliation* and the *goal of maintaining a positive self-concept*.

- With respect to compliance, accuracy is considered as a motive when people aim to accurately understand a request and respond appropriately, as this helps them to achieve a goal in an effective and maybe rewarding way. Affiliation is considered as a motive when people aim to establish and maintain social ties. Here approval and friendly cues help to create relationships as well as norms of social exchange. Furthermore, self-perception plays an important role in the context of compliance as individuals aim to behave consistent with their beliefs and self-ascribed characteristics.

- With respect to conformity, literature traditionally distinguishes informational and normative conformity. The former is motivated by accuracy (i.e., the desire to interpret reality correctly and behave in the right ways), the latter by affiliation (i.e., obtaining social approval from others). In [CG04] this differentiation is maintained. Additionally, the goal of maintaining a positive self-concept is considered as a third underlying motive as individuals often adapt the behaviors of others to either enhance or protect their self-concept.

The described typology is related to the earlier work [CT98], where three major components of inter-personal influence are distinguished, i.e., compliance, conformity and social norms, whereas later in [CG04], social norms are considered as mechanisms that can lead to both compliance and conformity.

Other works exclude conformity mechanisms from social influence: Here conformity is characterized as pretending to have a certain opinion or behavior to get approval by others without really having that opinion or being convinced of the behavior. Social influence, on the other hand, refers to a real change in a person's beliefs or behavior, and this change is caused by others that are perceived as similar to the individual, as role models or as experts or caused by the majority of a relevant social group [Ras07].

2.1.2 Formalizing Social Influence

Going beyond the individual level, the approaches presented in this section focus on interactions between individuals, which enable social influence, as well as the attempt to formalize or to systematically study these interactions.

One early attempt to formalize social influence processes was introduced by [FRC59] when studying social power. Power is captured in terms of social influence, which in turn is defined through a change in behavior, opinions, and other *"aspects of the person's psychological field"* ([FRC59], p. 251). Social influence can be exerted by another person, a group of people, a norm, or a role. A formula is introduced that captures power as a result of two forces, one in the direction of the influence attempt and one in the opposite direction.

In [Ras07] five main areas of research are distinguished, where social influence is treated in a formal way:

1. *Minority influence in group settings*: Whereas previous studies use to focus on individuals who adopt the behavior or opinion of the majority, research is now more and more interested in capturing settings where a minority within a bigger group aims to change the majority.

2. *Research on persuasion*: Persuasion is in general characterized as a *"change in attitudes or beliefs based on information received from others, focuses on written or spoken messages sent from source to recipient"* ([Ras07], p. 4427). Two approaches are distinguished: the so-called elaboration likelihood model of persuasion and the so-called heuristic-systemic models. Both models distinguish two *routes to persuasion*, depending on how much and in what way a person is thinking about the received message; i.e., a direct route, where arguments are considered consciously, and an indirect route, where non-direct means and considerations are crucial. For both models it has been shown that persuasion happening through the more direct route is more persistent [PC86, Ras07].

3. *Dynamic social impact theory*: Within the dynamic theory of social impact, social structure is considered as the result of individuals that are dynamically influencing one another. As individuals are stronger influenced by people nearby, local subcultures, i.e., local patterns of values, beliefs and behavior, emerge. As a consequence, social attributes that are initially randomly distributed get clustered and correlated due dynamic processes. Thus, dynamic social impact theory sees culture and society as complex systems based on self-organizing dynamics [Lat96].

4. *Structural approaches to social influence*: These approaches study influence between individuals within a larger network. As these approaches are central to our work, we discuss them in more detail in this section.

5. *Social influence in expectation states theory*: This theory examines relative influence within groups based on inequalities of its members. Findings show that even in groups where all members have the same initial status, hierarchies emerge due to interactions. Based on these interactions, group members develop expectations about future tasks and individuals for whom higher expectations are held will be able to exert more influence. In this context, creation and characteristics of status play an important role [Ras07].

It is concluded that future work should integrate the aforementioned approaches and it is suggested that *"a general model of social influence will need to incorporate group structures, the characteristics of the individuals in those structures, and the distribution of characteristics into majority and minority components"* ([Ras07], p. 4429). Our work is clearly an attempt to integrate these aspects.

Structural Approaches

Structural approaches examine influence processes between pairs of individuals within larger networks. The structure enables social influence. As a consequence, the behavior or opinion of a person reflects the behavior or opinion of the *social frame of reference* of the person; i.e., those people that are considered by the individual as an *"appropriate standard"* to compare herself against ([Lee02], p. 26). Importantly, there are two different mechanisms of social influence, i.e., *communication* and *comparison* (see also Figure 1.1):

- *Communication:* This summarizes social influence processes where an individual directly interacts with her frame of reference. Studies show that the more intense the communication between an individual and her alters, the higher the likelihood that the individual gets influenced. Friends and other personal contacts are crucial for an individual to develop preferences, to adopt behaviors and to form an opinion. A high number of studies on social influence focus on communication processes and it is well accepted that this mechanism plays an important role in numerous settings [Lee02].

- *Comparison:* Social influence through comparison occurs if an individual is searching for her social identity. The individual compares herself to others that are perceived as similar. The individual considers the behavior of the others as the appropriate or correct behavior for a person like her or a person in her position [Lee02]. Studies show that in some settings comparison mechanisms are more important to predict social influence than communication mechanisms, e.g., in [BD82], where it is studied which journals are regarded as important within a group of scientific experts. In [Bur87] it is analyzed to what extent talking to colleagues (i.e., communication), on the one hand, and network position (i.e., comparison), on the other hand, are crucial for physicians when prescribing a new drug. Also here comparison processes are more important than communication. Comparison leads to competition and to fear of loosing of status.

In the context of social networks, a change in behavior, opinion, etc. due to a social influence process is also called *contagion*. Furthermore, in a network setting communication processes are also referred to as *cohesion* and comparison processes as *structural equivalence* [Bur87, Lee02].

It is well known that physical proximity may foster contagion. In [Bur87] it is argued that in the last decades mass media and new technologies brought a shift from physical proximity to social proximity. Now, cohesion and structural equivalence generalize the concept of physical proximity by changing the emphasis from the question *"whether people are adopting in ego's physical surroundings"* to *"who is adopting"* ([Bur87], p. 1289).

In [Lee02] it is discussed that there is a difference between similar behavior and similarity of beliefs. It is argued that behavior is not only determined by beliefs or attitudes but more generally by limitations and restrictions for the individual. If the individual changes her belief, she is not always able to also change the behavior. Also people who behave in a similar way, do not necessarily possess similar beliefs. It is concluded that typically *"communication yields similarity of beliefs, but not necessarily of behavior, whereas comparison leads to similarity in behavior, but not necessarily in underlying beliefs"* ([Lee02], p. 28).

Traditional mathematical models to capture social influence describes the behavior or the opinion of an individual as the weighted average of the behavior or opinion of her network connections. These models have been extended [Fri98, Ras07]; however, more complex approaches to capture influence in networks have been introduced [Lee02, RPE01]. We discuss some of these models in detail in Section 2.2.4.

2.2 Empirical Methods and Frameworks

Related to the three levels that are considered in our work (see Table 1.1), we introduce methods to draw empirical solutions in this section. Furthermore, we discuss important concepts, which we need in the following analysis, and their formal description.

2.2.1 Regression Models

At the individual level, we focus on regression models in our analysis. As discussed previously, at this level we consider all individuals independent. They are described by different characteristics with the help of attributes or variables that can represent non-changeable characteristics such as gender or age as well as changeable ones such as an opinion or a certain behavior.

Regression is one central statistical approach to examine the relationship between a *response variable*, which might for example refer to the behavior of a person, and one or more *covariate variables*, which might refer to other characteristics of the same person. The response variable is also called *dependent* or *outcome variable*, and the covariates are also called *independent variables*, *predictor variables* or *features*.

We now assume that we have n observations and k covariate variables, i.e., the data is of the form

$$(Y_1, X_1), \ldots, (Y_i, X_i), \ldots, (Y_n, X_n) \tag{2.1}$$

where

$$X_i = (X_{i1}, \ldots, X_{ik}). \tag{2.2}$$

Here, Y_i is the response value of the i^{th} observation and X_i the vector comprising the respective k covariate values. All values are assumed to be real numbers. Then, the *Linear Regression Model* is defined as

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon_i \tag{2.3}$$

for the observations $i = 1, \ldots, n$, where ϵ_i is $\mathbb{E}(\epsilon_i | X_{i1}, \ldots, X_{ik}) = 0$, i.e., conditioned on the coviariates, the errors have a mean of zero. Typically, an intercept is included in the model by setting $X_{i1} = 1$ for all observations. The unknown parameters β_j are the regression coefficients. We can also write equation (2.3) in matrix notation

$$Y = X\beta + \epsilon, \tag{2.4}$$

where Y and ϵ are $(n \times 1)$ vectors, X is a $(n \times k)$ matrix, and β a $(k \times 1)$ vector. The regression coefficients are usually estimated with the help of the ordinary least squares (OLS) approach ([Was13], pp. 216-217).

Now we assume that Y is binary, i.e., $Y_i \in \{0, 1\}$ for $i = 1, \ldots, n$. Then, the *Logistic Regression* (*Logit*) *Model* is defined as

$$p_i \equiv p_i(\beta) \equiv \Pr(Y_i = 1 | X = x) = \frac{\mathrm{e}^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}}{1 + \mathrm{e}^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}}, \tag{2.5}$$

or

$$\mathrm{logit}(p_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}, \tag{2.6}$$

where

$$\text{logit}(p_i) = \log\Big(\frac{p_i}{1-p_i}\Big) = \log(\text{odds}(p_i)) \tag{2.7}$$

([Was13], pp. 223-224). To estimate the coefficients, typically maximum likelihood estimation is used. We see that the coefficients β_j in the logistic regression model, are log-odds ratios. The odd ratio captures the ratio of the probability p_i that a particular outcome will occur given a particular predictor compared to the probability $1 - p_i$ that this outcome will not occur given that predictor. Thus, the coefficient β_m represents the log-odd ratio between Y_m and x_{im} when all the other x_{ij}, $j \neq m$ are fixed. The regression coefficient β_m is the estimated increase in the log-odds of the outcome variable per one unit increase of the value of the predictor variable [Was13].

Models that are applied to data, which do not fulfill certain requirements are called mis-specified. In the context of regression, requirements are for example that the observations are independent and that data and errors are correctly distributed. There are some issues with mis-specified models, in particular they can lead to wrong estimations based on unreliable standard errors [FCS10, Sha16].

In our work we use R software including various packages to estimate regression models [R F16].

2.2.2 Geometric Data Analysis (GDA)

At the group level, we apply the approach of *geometric data analysis* (*GDA*) to capture social context [LRR04]. In general, GDA refers to those statistical approaches that model multivariate data, i.e., data that comprises of a number of different variables, as clouds of points in a metric space. The interpretation of the data is based on the interpretation of this representation. The GDA approach is considered to be *"geometric, formal and description-oriented"* ([LRR04], p. 6):

- *Geometric modeling:* One of the most important characteristics of GDA is its aim to represent the data that is studied in a metric space. Starting from the data, the first step is their metric specification, i.e., with the help of certain procedures, this data is transformed into clouds of points. These procedures include correspondence analysis, principal component analysis and multiple correspondence analysis. Furthermore, this metric specification allows to assign a distance between pairs of points. After this, the typically high dimensional clouds are projected onto lower dimensional sub-spaces, e.g., axes or planes, to facilitate a better interpretation.
- *Formal approach:* As GDA is well-founded in abstract linear algebra; mathematical theory is thus central to it. In particular, it is crucial for the steps described above, i.e., the construction of the clouds of points and the metric specification, as well as the spectral decomposition, i.e., the geometric projection onto a principal subspace.
- *Description oriented:* The GDA aims to foster an exploratory approach to the data under consideration with a focus on descriptive statistics. However, it does not

exclude statistical inference. In particular, these procedures can help to interpret the results.

Thus, GDA can be regarded as a research framework or a frame model that *"summarizes relevant knowledge, guides the collection of data and the interpretation of results"* ([LRR04], p. 14). In contrast to regression models (see Section 2.2.1), where the *relevant knowledge* is represented by the individual attributes or variables, this knowledge is summarized here by geometric means. This reflects two different concepts, i.e., *"sociology of variables"* on the one hand, and *"constructing a social space"* on the other hand ([LRR04], p. 14).

This framework can be leveraged to study the social space as it enables detailed analyses by focusing on the individuals and their position in this social space. Individual attributes lead to sub-clouds which allows for specific analyses. Furthermore, as soon as the social space is constructed new inquiries can be made. In particular, two aspects can be considered [LRR04]:

1. It can be examined whether external factors such as gender or age are able to explain the position of the individuals in the social space. Here, external factors refer to characteristics that have not been used when constructing the space.
2. Furthermore, it can be studied whether the position of the individuals in the social space indicates the opinion of the individual on certain issues.

With respect to our analysis we can phrase these aspects as: (1) *Can individual attributes explain the position of the individual in the social space?* (2) *Does the position in the social space allow to draw conclusions about the behavior of the individual?*.

When analyzing the clouds of points, one does not have to rely on visual clustering only as efficient non-visual clustering methods exist. Here, in particular, methods based on Euclidean distance are appropriate [LRR04]. One clustering approach is *k-means clustering*. With the help of this approach the points are assigned to k different groups or clusters, where the number of clusters k has to be fixed beforehand and provided to the algorithm. Then the algorithm chooses k cluster means randomly and assigns each data point to the cluster mean that is closest to this point with respect to chosen distance function, i.e., the Euclidean distance in our case. Thus, k clusters emerge. In the next step, the cluster means are re-calculated based on the points belonging to the cluster. Then, again, the points are assigned to the mean that is closest. This procedure is repeated iteratively until the cluster means do not change any more [Mar14]. There are several approaches to determine an appropriate number of clusters. However, as this is an explorative method it is also crucial whether the resulting clusters are meaningful in particular when studying social systems.

Overall, GDA is well suited for analyzing large-scale data, as the applied methods are apt to detect and reveal structure of highly complex data. Due to powerful computers the manipulation of high-rank matrices is typically not an issue either.

In the social sciences, GDA is a well-established framework. In particular, Bourdieu introduced and advanced this approach within his studies of lifestyle behavior. He shows that the position of an individual in the social space strongly impacts her lifestyle behavior and her preferences. This position determines the class the person belongs to and can be characterized by three forms of capital, i.e., the economic capital, the cultural capital, and social capital. The class of a person strongly influences cultural choices, which in turn reinforces class differentiation. People internalize this structure while building their identity. Thus, the differentiation of the social space into classes is maintained and reproduced [Bou84, SSW06]. Related work on lifestyle applying a similar approach shows that self-identification and cultural preferences are also strongly influenced by ethnic background, gender, and age of a person [KG99].

Today, these approaches are increasingly applied in the context of customer segmentation [TC11, Cho15]. Here the goal is to determine groups in which certain lifestyles, preferences and behavioral patterns are predominant in order to target these segments. Thus, here, social influence processes play a role on at least two levels, i.e., social norms that determine the behavior of customer segments as well as persuasion mechanisms that a company wants to apply.

Dimensionality Reduction

We now discuss two methods that enable geometric modeling of multivariate data in more detail, namely *principal component analysis* and *factor analysis*. The summary that we give is based on [Mar14, Sha16]. The idea of these methods is to find coordinate axes (also called latent dimensions) that capture the data in a way that some of the original dimensions are not required any more. Thus, these methods are used to reveal latent structure.

Principal component analysis is a technique that uses dependencies between variables to project the data into a lower dimensional subspace. The idea is to summerize m-dimensional data by projecting it onto a k-dimensional subspace. The q directions spanning the subspace are called principal components. These projections can be derived by identifying those projections that maximize the variance.

Principal components are directions in the data that have the largest variance, i.e., the first principal component is the direction with the largest variance, the second principal component is the direction with the largest variance among all other directions that are orthogonal to the first principal component, and so on. The idea is that, as the dimensions have less and less variance, some of them might not be required as they do not have high variability. On the contrary, often the results are even improved because some of the noise in the data is removed.

The PCA procedure works as follows: We assume that we have n data points $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)})$. First, the data is centered by subtracting the mean $\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_j^{(i)}$ of each column, i.e., $x_j'^{(i)} = x_j^{(i)} - \mu_j$.

Now, we define a $n \times m$ data matrix $\mathbf{X}$ containing the data points as row vectors. Then, matrix $\mathbf{X}$ should be rotated in order to place the data along the directions that maximize their variation. This is achieved by computing the $m \times m$ covariance matrix of $\mathbf{X}$

$$C = \text{cov}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{m}(x'^{(i)})^T(x'^{(i)}) \tag{2.8}$$

and its eigenvalues and eigenvectors

$$\mathbf{V}^{-1}\mathbf{C}\mathbf{V} = \mathbf{D}, \tag{2.9}$$

where the columns of matrix $\mathbf{V}$ comprise the eigenvectors of $\mathbf{C}$ and $\mathbf{D}$ is a $m \times m$ diagonal matrix that contains the eigenvalues of $\mathbf{C}$.

This is based on the following: To rotate the data, $\mathbf{X}$ is multiplied by a rotation matrix $\mathbf{P}$, which should be chosen in a way that the covariance matrix of the rotated data $\mathbf{Y} = \mathbf{P}^T\mathbf{X}$ is a diagonal matrix, i.e.,

$$\text{cov}(\mathbf{Y}) = \text{cov}(\mathbf{P}^T\mathbf{X}) = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & 0 & \lambda_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \lambda_n \end{pmatrix}. \tag{2.10}$$

Then it can be shown that

$$\text{cov}(\mathbf{Y}) = \mathbf{P}^T\text{cov}(\mathbf{X})\mathbf{P} \tag{2.11}$$

and, using the fact that $\mathbf{P}^T = \mathbf{P}^{-1}$, that

$$\mathbf{P}\text{cov}(\mathbf{Y}) = \text{cov}(\mathbf{X})\mathbf{P}. \tag{2.12}$$

By taking into account that $\text{cov}(\mathbf{Y})$ is diagonal and by setting $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_n)^T$ and $\mathbf{C} = \text{cov}(\mathbf{X})$ we obtain

$$\boldsymbol{\lambda}\mathbf{P} = \mathbf{C}\mathbf{P}. \tag{2.13}$$

Thus, $\mathbf{V} = \mathbf{P}$.

Now, to map the data onto a k-dimensional subspace, we sort the columns of $\mathbf{D}$ so that the eigenvalues λ_i are decreasing and apply the same order to the columns of the eigenvector matrix $\mathbf{V} = (v^{(1)}; v^{(2)}; \dots; v^{(m)})$

To retain k dimensions, we just set $\mathbf{V}_{red} = (v^{(1)}; v^{(2)}; \dots; v^{(k)})$, where $\mathbf{V}_{red}$ is a $n \times k$ matrix.

The new components $z^{(i)}$ can be determined as $z^{(i)} = x^{(i)}\mathbf{V}_{red}$.

One way to choose k is to consider

$$D = \mathbf{V}^{-1}\mathbf{C}\mathbf{V} = \mathbf{P}^T\text{cov}(\mathbf{X})\mathbf{P} = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & 0 & \lambda_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \lambda_n \end{pmatrix},$$

assuming that the eigenvalues λ_i have decreasing order ([Mar14], pp. 227-229). For a given k the variance that is retained can be calculated by

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{2.14}$$

For example, if we want to retain a variance of 80%, we can choose the smallest k for which

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \geq 0.80.$$

One "rule of thumb" is to keep all dimensions that have an eigenvalue bigger than or equal to one; the so-called scree plots can help to decide how many dimensions to keep. Here the idea is to visually inspect the eigenvalues in decreasing order. It is suggested to keep those dimensions until the plot levels off to the right ("elbow").

Another method for dimension reduction is *factor analysis*. Here the idea is to ask whether the observed data can be captured by a smaller number of latent dimension or uncorrelated factors. This approach is typically applied in psychology and other social science disciplines. One example that is often mentioned is related to IQ tests, where the outcome might be explained by a number of latent factors that are not directly observed.

Again we assume that we have n data points $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)})$, where we center the data by subtracting off the mean $\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_j^{(i)}$ of each column, i.e., $x_j'^{(i)} = x_j^{(i)} - \mu_j$, $j = 1, \ldots, m$.

Furthermore, we assume that the model that we are looking for has the following form:

$$\mathbf{X} = \mathbf{W}\mathbf{Y} + \boldsymbol{\epsilon}. \tag{2.15}$$

Here, $\mathbf{X}$ comprises the observations and $\boldsymbol{\epsilon}$ is the error of approximation or noise. As we expect that the latent factors $\mathbf{b}_i$ we are interested in are independent, we assume $\mathrm{cov}(\mathbf{b}_i, \mathbf{b}_j) = 0$ for $i \neq j$. Furthermore, it is assumed that the noise is normally distributed with a mean equal to zero and a variance $\boldsymbol{\Psi}$ with $\Psi_i = \mathrm{var}(\epsilon_i)$ for each element. Finally, it is also assumed that the noise variables are uncorrelated.

Next, we can write the covariate matrix $\mathbf{C}$ of the observed data as

$$C = \mathrm{cov}(\mathbf{X}) = \mathrm{cov}(\mathbf{W}\mathbf{b} + \boldsymbol{\epsilon}) = \mathbf{W}\mathrm{cov}(\mathbf{b})\mathbf{W}^T + \mathrm{cov}(\boldsymbol{\epsilon}) = \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}, \tag{2.16}$$

where we use that $\mathrm{cov}(\mathbf{b}) = \mathbf{I}$, since the latent factors are considered as uncorrelated. The goal of the factor analysis procedure is now to identify factor loadings $\mathbf{W}_{ij}$ and values for the noise $\boldsymbol{\Psi}$ to reconstruct $\mathbf{X}$ or to approximate $\mathbf{X}$. This is typically done with maximum log-likelihood estimation ([Mar14], pp. 234-236).

The results of the factor analysis are often rotated to make the output better interpretable. This can be done by a typically orthogonal transformation $\mathbf{o}$ of the factor loadings

$\mathbf{W}' = \mathbf{Wo}$ and scores of the latent variables $\mathbf{b}' = \mathbf{o}^T\mathbf{b}$ but without changing the observed data $\mathbf{X}$ at all ([Sha16], p. 400):

$$\begin{aligned} \mathbf{X} &= \mathbf{Wb} + \boldsymbol{\epsilon} \\ &= \mathbf{Woo}^T\mathbf{b} + \boldsymbol{\epsilon} \\ &= \mathbf{W}'\mathbf{b}' + \boldsymbol{\epsilon}. \end{aligned}$$

For example, varimax rotation is a commonly used rotation method. Here the rotation aims to maximize the variance of the squared loadings of one factor on all variables in the factor matrix (see [Abd03]).

In order to determine the number of factors to extract, "rules of thumb" also exist in factor analysis, e.g., based on the percentage of the variance that should be explained, based on the eigenvalues larger than or equal to one or based on the "elbow" in the scree plot.

For the statistical analysis conducted at the group level, we use R software [R F16] including various R packages for factor and cluster analysis.

2.2.3 Social Network Analysis

Social network analysis has a long tradition in the social and behavioral sciences. In the focus of this research approach are relationships between individuals or other interacting entities (e.g., groups or organizations) and the implications of these relationships. Besides its emphasis on relational concepts and processes, the following aspects are crucial for the social network perspective [WF94]:

- The individuals or entities and their actions are considered as interdependent rather than independent;
- Relations between the individuals facilitate the transfer of material as well as immaterial resources (e.g., innovation, ideas or behavior);
- Network models see individual actions influenced by the position of the individual within the network structure. This structure provides opportunities for the individuals but also constraints them; and
- Social, economical or political structure is conceptualized by network models *"as lasting patterns of relations among actors"* ([WF94], p. 4).

Conventional approaches in the social sciences typically ignore relations. As discussed previously, such approaches focus on attributes of independent individuals. In empirical studies often random samples of a larger population are taken, then a variety of individual attributes are measured, and associations between those attributes are modeled. However,

if the goal is to analyze human behavior it might be wrong to consider the individuals as independent. The social network perspective, on the other hand, understands individual behavior and other individual attributes in term of the relationships of that individual. The individual position within the network structure is considered primary, whereas attributes of individuals are considered secondary. Furthermore, obviously the relation between two entities is a property of this pair and thus goes beyond individual characteristics. Therefore, when analyzing network data, measurements on the relations are required. Furthermore, network analysis techniques should be applied rather than standard statistical methods such as regression, t-tests and correlations, as those methods typically cannot be used in this context (they require the data to be independet) [WF94].

Overall, social network analysis can be seen as a research framework that integrates social science theories, empirical methods as well as mathematical descriptions. In recent years it has also become increasingly popular in disciplines other than social science, in particular in computer science. The reason is the tremendous amount of static and dynamic relational data that is available today, coming from the Web and other sources, making the design and development of large-scale computational, mathematical and statistical techniques inevitable. As these challenges are inherently interdisciplinary, new fields such as computational social sciences emerge [THC+15]. Furthermore, the term *Web Science* has been established referring to an interdisciplinary research field that aims to study and design the Web at different levels, i.e., the micro level (i.e., small technological innovations) as well as the macro level (i.e., large scale phenomena that affect society and commerce globally) [HSH+08].

In [New10] a comprehensive overview of state of the art concepts and methods for analyzing static networks and dynamic processes on networks is presented. In [EK10] it is shown that networks and the ability to analyze them play a crucial role in modern society.

Basic Concepts and Definitions

Before we discuss social influence in this context, we introduce some basic concepts of network analysis and give some relevant definition (see [PBNW11, New10]). The formal description is not restricted to social networks but applies to general networks.

A network is formally described by a graph $G = (V, E)$. The set V is called nodes (or vertices), and the set E consist of edges (or links or ties) between pairs of nodes. Two nodes that are connected by an edge are called neighbors. If there is an edge between each pair of nodes, G is called a complete graph. In a directed graph, each edge has an origin and a destination (capturing asymmetric relations). As opposed to this, an undirected graph comprises edges with no orientation (capturing symmetric relations). To incorporate additional information, labels can be assigned to nodes to capture variables on the node level (e.g., individual behavior) and/or the edges of a graph to capture characteristics of the dyadic relationship (e.g., the strength of the relationship).

The degree $\deg(v)$ of a node v in an undirected graph G is the number of neighbors v has. The average degree of graph G is the arithmetic mean of all degrees $\deg(v_i)$, $v_i \in V$. Obviously, in a directed graph, the in-degree and out-degree of a node are separately considered: in-links are connections pointing to a node, out-links are those pointing to some other node.

A path in a graph is a sequence of nodes such that two consecutive nodes are connected by an edge. The number of all such edges is called the length of the path. The distance (or geodesic distance) $\mathrm{d}(v, w)$ between two nodes v and w in a graph is defined as the length of the shortest path between them. The average distance in a graph is the arithmetic mean of the distances between all pairs of nodes. A path with at least three edges is called a cycle if the first and the last nodes are the same, but otherwise all nodes are distinct. If there is a path from a node v to a node w, these nodes are said to be connected. A connected component in a graph is a set of nodes in which a path exists between any two nodes. The diameter of a graph is defined as the longest possible distance existing in the network, i.e., the maximal distance between any two connected nodes. When considering directed graphs, the edges' orientations have to be taken into account, and the respective definitions are adapted accordingly. In this case two connected components are defined, i.e., strongly and weakly connected components. In a strongly connected component (SCC) there is a directed path from each node to any other. In a weakly connected component there is also one path from each node to any other, but the edges' orientation is ignored.

Now, let G be an undirected graph and n the number of its nodes, then the density ρ of G is defined as the number m of edges of G divided by the maximum possible number of edges (i.e., those present if G were a complete graph):

$$\rho = \frac{2m}{n(n-1)}. \tag{2.17}$$

With the help of the local clustering coefficient, local density can be captured. Let k_i be the number of neighbors of a node v and e_i the sum of all edges between them. If each pair of neighbors of node v were connected by an edge, then there would be a number of $\frac{k_i(k_i-1)}{2}$ edges. Therefore, the clustering coefficient C_i of a node v is:

$$C_i = \frac{2e_i}{k_i(k_i-1)}. \tag{2.18}$$

Hence, C_i reflects the probability that two arbitrary selected neighbors of v are connected by an edge. Looking at directed graphs, there can be two edges between each pair of nodes, i.e., one in each direction. Taking this into account, both the formula for the the density ρ and the clustering coefficient C_i have to be divided by two.

The level of clustering of the entire network can be quantified by the global clustering coefficient C:

$$C = \frac{3 \cdot (\text{Number of closed triangles})}{(\text{Number of connected triples})}. \tag{2.19}$$

Here, *connected triple* refers to three nodes u, v and w with one edge connecting u and v and one edge connecting v and w; the edge connecting u and w can be present or not. Hence, a closed triangle contains three distinct connected triples and thus it contributes three to the number of connected triples. Usually, this formula is applied to both undirected as well as directed networks. There is an alternative definition to capture the overall clustering level of the network, where the clustering coefficient C' of the entire graph G is defined as the arithmetic mean of the clustering coefficients C_i of all nodes (v_i), $v_i \in V$.

As the network perspective should provide a better understanding of real-world structures, some concepts have been proposed to facilitate a richer interpretation. In this context, a very important category is the class of the so-called centrality indices. They try to formalize the idea that in many settings some nodes or edges might play a more important role than others, hence they should be considered as more central. The three basic centrality indices are: degree centrality C_D, closeness centrality C_C and betweenness centrality C_B.

The degree centrality $C_D(v)$ of node v is defined as the number of edges it is connected to:

$$C_D(v) = \deg(v). \tag{2.20}$$

In a directed graph two kinds of degree centrality are usually distinguished, namely in-degree and out-degree centrality. The closeness centrality $C_C(v)$ of node v is defined as the reciprocal value of the sum of all distances between v and each other node w in the network:

$$C_C(v) = \frac{1}{\sum_{w \in V \setminus v} \mathrm{d}(v, w)}. \tag{2.21}$$

The betweenness centrality C_B for node v is defined as:

$$C_B(v) = \sum_{u \neq w \neq v \in V} \frac{\sigma_{vw}(v)}{\sigma_{vw}}. \tag{2.22}$$

Here σ_{uw} denotes the number of shortest paths between node u and node w and $\sigma_{uw}(v)$ the number of shortest path between those nodes that run through v. Usually the values calculated for these indices are normalized. When considering directed networks, the centrality indices are typically calculated separately for in- and out-links. Furthermore, for the closeness and the betweenness centrality also slightly different definitions are common. Another well-known centrality index for directed networks is PageRank PR [BP12]. The PR of a node v is defined as

$$\mathrm{PR}(v) = \frac{1-s}{n} + s \cdot \sum_{w:\mathrm{d}_{Out}(w,v)=1} \frac{\mathrm{PR}(w)}{\deg_{Out}(w)}. \tag{2.23}$$

The last term considers all nodes w that point at node v, i.e., in the directed network their out-distance d_{Out} to v equals one. Here, n denotes the number of nodes in the network, $\deg_{Out}(w)$ the out-degree of a node w, and s is a damping factor. This recursive

formula expresses that the PageRank of a node not only depends on the number of other nodes pointing at it but also on their PageRank. This centrality index has been introduced in the context of Websites. It is a so-called eigenvalue centrality index because the PageRank of the nodes convert to the eigenvalues of a certain matrix [EK10].

There are a number of properties that have been discovered in many networks that represent "real-world" phenomena. One of them is the so-called power-law degree distribution; i.e., the degree distribution of the nodes in the network can be approximated by a function of the form $p(k) = c \cdot k^{-\gamma}$, where $k = \deg(v)$ denotes the degree of a node v and $c \in \mathbb{R}$ and $\gamma \in \mathbb{R}$ are positive constants. This implies that in such a network the majority of nodes have a very low degree while very few nodes have an extremely high degree, thus they are acting as hubs in the network. One important mechanisms for obtaining such a topology has been discovered in the fact that links are not added randomly but are attached to specific nodes preferentially. Such networks are called scale-free.

The so-called small-world property is another common property of networks expressing that the average distance within the network is relatively short. The term refers to an experiment conducted by Stanley Milgram in 1967 with the aim to study the average distance of social networks of people in the US [EK10]. Furthermore, many "real world" networks exhibit a community structure; i.e., the nodes of the network can be divided into groups within which the edges are denser than between different communities. Another property of many networks is the presence of a very large connected component, also called giant component, which contains the vast majority of all nodes.

Studying Social Influence in Networks

In social networks, the phenomenon that individuals are prevalently connected to others with same characteristics is often observed empirically. This is typically caused by various mechanisms that are intertwined when the network is shaped: *social influence* refers to the change of a person's behavior that is affected by other individuals in the network. As compared to this, *social selection* occurs when relations in the network tend to be formed between individuals with the same attributes. *Covariate effects* are mechanisms that refer to the adaption of a person's behavior due to other, maybe unknown factors [ST11]. To see the differences, let's consider the following example: In an online social network, such as Facebook, there are large groups of users who are all friends and support the same sports team. When looking at this setting at one point in time, we cannot say in beforehand whether those users first were friends and then, because some of them were committed to supporting the sports team, the others started to do so as well (i.e., social influence). However, maybe they rather became friends because they have, as fans of the same sports team, shared interests (i.e., social selection). Finally, there might also be other reasons why they all became friends and why they are all supporting the same team, maybe they are all living in the same area (i.e., covariate effects). What we observe empirically will usually result from all of those mechanisms.

To gain better insights into those effects and how they work, it is important to capture them separately. This enables us to address distinct phenomena: an understanding of social influence helps to model the diffusion of a new behavior through a network, an understanding of social selection helps to model the emergence of structural properties within a network. In [CCH+08] this difference is exemplified with the help of viral marketing and recommender systems; the first is build on the idea that social ties can serve as predictors for future behavior, whereas the latter delivers predictions based on social similarity.

The dependencies addressed above are considerably complex. Thus, to address them in an adequate way, advanced statistical, mathematical and computational methods are needed. Existing models that are capable of taking all the dynamics into account, i.e., social influence and social selection mechanism as well as covariate effects, are typically temporal models [Sni11, CCH+08]. Such models make use of information gained from distinct observations (i.e., different time steps) to model or predict future effects. The most important models in this context are SIENA models [SSP10]. These longitudinal actor-based models are capable of addressing both social influence and network formation over time. Furthermore, they are geared towards statistical inference and there is an R package to fit such models [Tom15b], which are computational very expensive.

If only cross-sectional data should be considered (i.e., data that is aggregated or that stems from one observation of the network), this task becomes even more challenging. In general, statistical models for cross-sectional data either focus on social selection processes or on social influence processes. In the first case, all nodal attributes are fixed to infer network formation processes (e.g., which links are likely to form); and in the latter case the network structure is fixed to infer how behavior might be related to the position of an individual with respect to the behavior of others [REP01, RPE01].

When studying cross-sectional models for social selection, the so-called Exponential Random Graph Models (ERGMs) are widely applied [LKR12, MC03]. ERGMs are statistical models that allow inference on link forming processes of networks. The main idea is to assign a probability to a given network. This probability is derived by comparing the propensity of the structure of the network to the propensity that would occur only by chance [RSW+07, SPRH06].

In this work we focus on models for cross-sectional network data with respect to their ability to capture and quantify social influence. In this context, two important types of models exist: In [RPE01] a generalization of ERGMs is proposed to model social influence processes. These models for binary behavior are called Autologistic Actor Attribute Models (ALAAMs) [DR13]. If the studied behavior is represented by a continuous variable, auto-regression approaches such as Linear Network Autocorrelation Models (LNAMs) are appropriate [Lee02]. These models are related to spatial regression methods. They can be considered as extensions of ordinary least squares regression for networks as they can incorporate covariate effects as well as network structure [Dor89, Lee02]. They are also called *Network Effects Models* and are based on earlier work on spatial autocorrelation [Ans88]. They are less complex than ALAAMs but a weighted

matrices representing the strength of the relations between the individuals has to be provided. We will discuss ALAAMs and LNAMs below in more detail as we will use these approaches for our analysis. Furthermore, *Conditional Random Fields* (CRF) *Models* as advanced machine learning methods to predict network data will also be presented [LMP01]. We propose this approach to capture social influence processes in social networks.

Apart from these approaches, there is also a branch of research that studies social influence in networks with respect to influence maximization. Here, the goal is to maximize the adoption of a product or the spread of an opinion by identifying appropriate seed users. Typically diffusion models and other computational models are used [KKT03]. However, statistical inference is usually not possible in such models. There is also research that aims at identifying influential users in online discussion forums. Here, typically users with high network centrality measures such as PageRank are considered as influential. Forum threads are used to derive user interaction networks as a basis for the analysis [ZAA07].

A collection of articles on data analytics in the context of networks is presented in [Agg11]. Here, a variety of topics is addressed and different applications are discussed. One article in this collection is a thorough survey of algorithms for social influence analysis in the context of networks [ST11]. Topics that are discussed in more detail are influence and social similarity, influence-related statistics and influence maximization in viral marketing. However, mainly temporal models are introduced, many of them with the goal to rank the nodes according to their ability to influence others, rather than to understand the influence process itself. Another survey within this collection addresses node classification [BCM11]. This topic also is related to social influence as certain problems in this context can also be phrased as label-propagation problems. Two approaches are presented: the first one is related to iterative applications of standard classifiers that are able to include nodal attributes, while the second approach models label-propagation with the help of random walks.

Most of the social influence studies in the context of networks focus on communication based processes, which is also the focus of our work. In order to capture comparison mechanisms the concept of equivalence with respect to the network structure plays a major role. Here typically three different types of equivalence are distinguished, i.e., structural equivalence, automorphic equivalence and regular equivalence [HR05, WF94]:

- *Structural equivalence:* Two nodes are called structurally equivalent if they have exactly the same links to the exactly the same other nodes in the network. This is the most restrictive form of the three types of equivalence discussed here. However, there are ways to relax this definition in order to capture to what extend two nodes are structural equivalent [Lee02].

- *Automorphic equivalence:* Two nodes are called automorphically equivalent if they have *"indistinguishable structural locations in a network"* ([WF94], p. 469), i.e., the nodes are embedded in the same way in the network structure and having

identical links to nodes who are themselves automorphically equivalent. If all labels in a network are removed, nodes that are automorphically equivalent are indistinguishable. Mathematically spoken, we can find some automorphism φ that maps one node onto the other.

- *Regular equivalence*: This form of equivalence refers to social roles within a structure. Two nodes are called regularly equivalent if they are linked to other nodes that are in the same social position as well. Thus, they do not have to be linked to identical nodes but to the same type of nodes, nor the number of links is crucial. For example, two professors can be considered as regularly equivalent. They have connections to PhD students, master students, administrative staff, i.e., to people who occupy the same social roles. In this perspective, it does not matter how many students the professors have. Instead, the social role results from the social context, which is captured within the social network approach by structural relation.

There are different approaches to determine equivalent nodes in networks. However, in general this is quite complex [HR05, WF94]. In [Lee02] it is discussed how comparison-based social influence can be operationalized in the context of LNAMs using a relaxed form of structural equivalence. In [Bur87] a structural equivalence model is discussed. It is emphasized that in these models the social frame of reference shifts from the neighbors of an individual to the entire social system. Furthermore, it is stated that communication based social influence gets replaced by competition and status. Structural equivalence is defined in its most restrictive form. This is also the case in [FJ97], where the adoption of an opinion in the context of structural social positions is discussed. Here a dynamic model is presented showing that the resulting behavior of an individual is potentially a very complex consequence of social position and interpersonal agreements. In [Fri01] a mechanism that combines norm formation and network theory is introduced, where both social position and interpersonal influences are crucial in the norm formation process.

2.2.4 Social Influence Models

When studying social influence phenomena in networks, the ideal scenario would be that detailed observations of network structure, behavior and individual attributes are available. However, in these cases SIENA models (see above) can be applied to separate social influence from other mechanisms, e.g., social selection. However, in many empirical studies with the aim to characterize individual behavior (e.g. political opinion), to model individual decision processes, and to predict future behavior, only one large network is given, e.g., Twitter following relations. Thus, cross-sectional models are required. These models have to be capable of explicitly accounting for interdependencies among the nodes. However, not many models exist in this context. We now discuss the most important ones, i.e., LNAMS and ALAAMs. Furthermore, we introduce CRF Models as a novel approach to capture social influence in networks. These models also form basis for the analysis conducted in Chapter 4.

Linear Network Autocorrelation Model (LNAM)

LNAMs are defined by the equation

$$y = \rho W y + X\beta + \epsilon. \tag{2.24}$$

The vector y represents the outcome variable, i.e., the behavior of interest. However, y also appears on the right hand side of Equation 2.24 as a predictor variable. This captures the idea that the behavior of an individual is influenced by the behavior of all the other individuals she is connected to. Thus, the behavior is both the outcome and predictor variable at the same time. The weighted adjacency matrices of the network is represented by W. An adjacency matrix A is a matrix that captures the structure of the network; it comprises the elements a_{ij}, which equal one if there is an edge between node i and node j and zero otherwise (for undirected networks) [New10]. The scalar ρ is called network effects parameter or contagion parameter. Thus, the first term of Equation 2.24 captures social influence, i.e., the impact of the neighbors' behaviors on the behavior of an individual. Therefore, the influence process is modeled as a weighted linear combination of the neighbors' behaviors.

Matrix X contains further individual attributes (covariates) and vector β the corresponding parameters. Thus, the second term of Equation 2.24 captures the intrinsic opinion of an individual. The error terms are represented by ϵ and it is assumed that they are normally distributed, i.e., $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. For parameter estimation maximum likelihood is used. [Dor89, Lee02]. If there are no contagion effects, i.e., the first term is equal to zero, the model is equivalent to OLS regression (see Section 2.2.1). As discussed in [Lee02] the weighting scheme of matrix W is critical for the outcome. This has to be considered in the specification of W. For the analysis we use the R package *sna* [Car10], which contains a function *lnam* to fit LNAMs. The computational costs are medium high.

In [Lee02] also network disturbance is discussed. In their work, ϵ in Equation 2.24 is expressed as $\epsilon = \rho' W' \epsilon + \nu$. The idea is to incorporate disturbance network parameters that capture the mechanisms which influence the individual to deviate from her behavior. Furthermore, as discussed in Section 2.2.3, [Lee02] also aims to operationalize structural equivalence in the context of LNAMs. However, both aspects are not considered in our analysis.

Autologistic Actor Attribute Model (ALAAM)

The second type of models that we consider are so-called Autologistic Actor Attribute Models (ALAAMs). These models allow to account for social influence processes [RPE01, DR13]. In these models, a binary nodal attribute (e.g., whether or not a node exhibits a certain behavior) is the outcome variable. The structure of the network is considered as predictor variable. The following equation describes ALAAMs:

$$\Pr(\mathrm{Y} = \mathrm{y} | \mathrm{G} = \mathrm{g}, \mathrm{X} = \mathrm{x}) = \frac{1}{\kappa(\theta)} \cdot \mathrm{e}^{\sum \theta \cdot z(y,g,x)}, \tag{2.25}$$

where Y represents the binary vector capturing the behavior of interest and G the adjacency matrix of the observed network. The matrix X contains further individual attributes (covariates). Furthermore, θ is a parameter vector and z contains statistics of network-attributes configurations, including interactions of the dependent attribute y, network structure g, and other covariates x [RPE01, DR13].

Thus, the distribution of behavior is studied across network ties; the joint probability of network and behavior is modeled. The main idea is that the behavior of one node might depend on the behavior of other nodes. So, the model allows to gain insights into these interdependencies.

Equation 2.25 can also be written in the following form:

$$\begin{aligned}\text{logit}(\Pr(Y_i = 1|y_{-1}, g, x)) = \theta_0 + \sum \theta_p z_p(g) + \sum \theta_I z_I(g, y) \\ + \sum \theta_C z_C(x) + \sum \theta_{IC} z_{IC}(g, x).\end{aligned} \tag{2.26}$$

Similar to the Logit Model (see Equation 2.6), the parameters represent log-odds ratios. Here, Y_i is the behavior at node i and y_{-1} refers to the behavior of all other nodes $j \neq i$ in the network. The parameter θ_0 denotes an intercept term. The parameters θ_P assess the impact of the structural position of node i on its behavior Y_i; the variables z_P are related to the corresponding network configurations, e.g., they might refer to the degree of node i, to the number of triangles it is involved, etc. The parameters θ_I capture social influence. They assess the impact of constellations on the outcome behavior, in which neighbors of node i also display the behavior under consideration. The statistics z_I represent the corresponding configurations, e.g., z_I might refer to the number of pairs of nodes, where both display the outcome behavior. The parameters θ_C capture effects of the covariates z_C. Finally, there is also the possibility to test for influence of covariates of other nodes on the outcome behavior. This is expressed in the last term of Equation 2.26. A detailed overview of all configurations that can be considered in this type of model is given in [DR13].

If there are no structural effects, then $\theta_P = \theta_I = \theta_{IC} = 0$ and we get a logistic regression model comprising θ_0 and θ_C as parameters. In that sense ALAAMs can be seen as a generalization of logistic regression for networks.

For estimation *Markov Chain Monte Carlo* (MCMC) methods are used. This is very complex as the joint distribution of network and behavior is simulated. Thus, the computational costs are considerably high and there are scalability issues.

To estimate ALAAMs we use iPnet [WRP06], a java application that facilitates the estimation and simulation of ALAAMs. A binary outcome behavior can be modeled taken into account certain types of configurations. As described in Equation 2.26, these configurations can be related to network position effects (parameters θ_P), to network attribute effects (parameters θ_I) or to covariate effects (parameters θ_C and θ_{IC}).

In [FH12] a similar approach is presented. The cross-sectional models, which are introduced in this work, are called *Exponential-family Random Network Models* (ERNMs)

and aim to consider jointly links between nodes, as well as nodal attributes as endogenous variables. In consequence, they generalize both ERGMs and ALAAMs. However, although in [FH12] an implementation of this approach as an R package is announced, we could not find this package. Thus, we do not consider this approach in our analysis.

Conditional Random Field (CRF) Models

Another way to address interdependencies between nodes is facilitated by *Conditional Random Field* (CRF) *Models*. The definition of conditional random fields is given as follows:

Definition. *Let $G = (V, E)$ be a graph such that $Y = (Y_v)_{v \in V}$, so that Y is indexed by the vertices of G. Then (X, Y) is a conditional random field in case, when conditioned on X, the random variables Y_v obey the Markov property with respect to the graph: $\Pr(Y_v|X, Y_w, w \neq v) = \Pr(Y_v|X, Y_w, w \sim v)$, where $w \sim v$ means that w and v are neighbors in G* (see [LMP01], p. 5).

Furthermore, using the fundamental theorem of random fields, it can be shown that then the conditional distribution can be written as

$$\Pr(\mathrm{Y} = \mathrm{y}|\mathrm{X} = \mathrm{x}) = \frac{1}{Z(x)} \cdot \mathrm{e}^{\sum \lambda \cdot f(e,y) + \sum \mu \cdot g(v,y,x)} \tag{2.27}$$

(see [LMP01, Bak07]). Thus, the impact of the edge related parameters (i.e., first term in the sum) is independent of the node related parameters (i.e., second term in the sum). Here, $Z(x)$ is a normalizing function. If there is no network structure, only the second term in the sum remains. In that case the model reduces to the Logit Model when setting

$$Z(x) := 1 + \mathrm{e}^{\beta_0 + \sum \beta_j x_{ij}} \tag{2.28}$$

in Equation 2.5. Hence, CRF Models can be regarded as an extension of logistic regression to networks.

CRF Models are discriminative undirected probabilistic graphical models. Probabilistic graphical models are graph-based representations of the dependency structures of random variables [KF09]. If one cannot assign a directionality to this dependency structure in an obvious way, the probabilistic graphical model is undirected. Discriminative models only focus only on the conditional distribution of $\Pr(\mathrm{Y} = \mathrm{y}|\mathrm{X} = \mathrm{x})$ rather than studying the joint distribution $\Pr(\mathrm{Y} = \mathrm{y}, \mathrm{X} = \mathrm{x})$ as in the context of generative models. One important advantage is that for the discriminative model all dependencies, which only involve covariates, i.e., variables in X, do not have to be considered. This leads to a much simpler structure than in the generative case [SM11].

As we are interested in general graphs, i.e., graphs that contain cycles, exact inference is in general not feasible (exact inference refers to the calculation of the marginal probabilities and the normalization function Z). However, loopy belief propagation (LBP) has proven to work very well in practice [Bak07]).

Overall, CRF Models show a high flexibility and ability to learn complex dependencies and to perform inference in large scale settings with a high number of variables. Due to these advantages CRF Models are today widely applied to various domains including natural language processing, computer graphics and bioinformatics [SM11]. We propose to apply them to capture social influence in networks, which is clearly a new field of application. The approach has in particular the advantage that it can be applied to large scale networks as it is highly scalable, i.e., it can handle a high number of variables and nodes and is very flexible regarding the network structure. Furthermore, we can include parameters that capture not only the previously described contagion effect, but also the impact for any combination of behavior: while the contagion parameter accounts for edges that connect two nodes displaying the "new" behavior (i.e., $y_i = 1$ and $y_j = 1$), we can also consider parameters capturing the effect of edges that connect two nodes displaying the "old" behavior (i.e., $y_i = 0$ and $y_j = 0$) as well as parameters capturing combinations of "old" and "new" behavior (i.e., $y_i = 1$ and $y_j = 0$; and $y_i = 0$ and $y_j = 1$ respectively). This is not possible in ALAAMs nor in LNAMs. In consequence, CRF models have a high practical value compared to the other models that do not scale well.

Although CRF Models are widely applied there is not much previous literature dealing with these models nor are there many software libraries available. Some software is available for linear chain CRFs as they are frequently used in the context of natural language processing. However, as we focus on general graphs, it does not serve our purposes. An exception is the software *UGM: Matlab code for undirected graphical models* [Sch16], which contains a number of functions for decoding, inference, sampling and training of undirected graphical models. In particular, CRFs on general graphs can be trained and a function for the inference method LBP is provided. For our analysis, a few adaptions were required. In particular, we extended the code to compute standard errors and p-values of the intercept and the nodal parameter μ. A further extension for the edge parameters λ is part of future work. In our analysis we used simulation techniques to estimate the significance of these parameters.

ALAAMs (and also ERNMs) are Markov Random Fields (MRFs), i.e., generative undirected probabilistic graphical models. CRFs are MRFs conditioned on a set of observed variables, i.e., a subcategory of MRFs. Hence, there is a relation between the introduced CRF Model and the previously discussed ALAAMs (and also ERNMs). However, the MRFs based models are more general and capture social influence in a more dynamic setting since the joint distribution of network and behavior is modeled. In our case, we focus on capturing social influence in a quite stable environment, i.e., the structure is fixed. In particular as this allows to apply such models to large scale data in an efficient way.

2.3 Application Domains

We now introduce the settings and domains that we will use for the empirical analysis. Again we organize the discussion based on the three levels of analysis (see also Figure 1.1).

2.3.1 Collective Behavior and Preferences

We start with an overview of recommender systems and touristic preference models (based on [NSSW14b], pp. 51-54) as tourism recommender systems are a major focus of our work.

Recommenders in Tourism

Recommender systems are defined as *"software tools and techniques providing suggestions for items to be of use to a user"* ([RRS11], p. 1). Such *"items"* are often books, movies or pieces of music, but the term can also refer to more complex products or services. There are several well-established techniques that are applied to predict whether an item is in fact useful to the user: With a content-based approach, recommended items have similar attributes as items that the user has liked before. By applying a collaborative filtering technique, items that have been liked by similar users are considered to be important. In this approach, the similarity of users is usually defined in terms of their past rating activities. Demographic systems recommend items based on demographic characteristics of a user such as age and gender. With a knowledge-based approach, domain knowledge about the preferences of a user regarding the attributes of the items is deployed for recommendation. By applying a community-based approach items are recommended to a user whose friends have liked such items before – for example in an online social network. Hybrid systems combine some of the previously mentioned techniques. As any recommender system aims to provide personalized suggestions, all described techniques rely on knowledge about the users. Thus, ever recommender system has to comprise a user model or user profile, where this knowledge is accumulated [JZFF10].

In [BR11] it is argued that suitable recommender techniques in the context of tourism usually are content-based and/or knowledge-based approaches. This conclusion is based on the characterization of tourism as a domain with high risk (i.e., the price of the items is comparatively high), low churn (i.e., the value or relevance of items does not change so rapidly), low heterogeneity (i.e., the needs that the items can satisfy are not so diverse), unstable preferences (i.e., user preferences in the past might not be valid anymore today) and explicit interaction style (i.e., a user needs to formulate an opinion or perform a search in order to add personal data). Furthermore, scrutability is required, which means that the reason for recommending an item should be transparent.

Thus, compared to books, Websites or movies, tourism related items present a considerably higher complexity and intangibility as it is discussed in [WK99]. To address different aspects of a proposed trip, a recommendation should include a bundle of distinct products such as attractions, accommodations and means of transport. Furthermore, this feature calls for a content or knowledge based approach. Content and knowledge respectively are needed to bundle certain features that might be different but supplementary; for example, to combine a hotel with a flight, both need to be available in the same region and period of time, etc.

In the area of tourism the number of ratings of a user is usually lower than in the movie or music domain. This is another problem when applying recommender techniques. As a consequence, the user profiles might be less accurate. Although those difficulties exist, tourism-related Websites incorporate recommender systems [WSZ+06]. Here two types of systems can be distinguished, namely systems that focus on destination selection and systems that recommend a bundle of activities to be performed at a certain destination [GSO11].

A major objective of our work is to facilitate the elicitation of user needs and preferences as we described in our studies [NSSW14a, NSSW14b]). A similar objective is followed by critique based recommender techniques [RN07, MR11]. However, these systems typically focus on the conversational process, where first results are iteratively refined. Users do not have to specify their preferences from the very beginning, but follow several cycles. Furthermore, some initial input (e.g., by answering some questions) or initial examples are required. In [RWZ05] initial examples are pictures of hotels, and the user is given the opportunity to iteratively explore the solution space in a graphical manner. Although this approach is comparable to our approach, there is a crucial difference: In [RWZ05], pictures represent products, whereas in our case pictures reflect individual user preferences.

Our approach has some similarity to preference construction as a model of human decision making [CdGF+13], to the extent that our users do not have a clear picture of their preferences from the very beginning. In these approaches preferences are typically constructed in a sequential interaction and decision process, whereas our users have to select all pictures in one step.

Touristic Preference Models

Since the 1970s research has tried to relate touristic behavioral patterns to psychological needs and expectations [Coh74, Pea82]. The work of Yiannakis and Gibson has a high impact in this context. In [YG92], a framework is developed to measure touristic role preferences. Fifteen pre-defined tourist roles, namely *Action Seeker*, *Active Sport Tourist*, *Anthropologist*, *Archaeologist*, *Drifter*, *Educational Tourist*, *Escapist*, *Explorer*, *High Class Tourist*, *Independent Mass Tourist*, *Jetsetter*, *Organized Mass Tourist*, *Seeker*, *Sun Lover* and *Thrill Seeker* are addressed by 30 questions in a questionnaire about touristic behavior, i.e., two questions per role. The second part of the questionnaire is about psychological needs and their satisfaction. There are several needs listed including the need for home and family, the need for control, the need for safety and personal security, and companionship needs. Additionally, demographic characteristics such as age and gender have to be provided. A further study based on this work is presented in [GY02]. Here the association between the tourist roles and psychological needs for both genders over lifetime is investigated. The results show that touristic behavioral patterns are related to psychological needs and that they change during time. Based on findings of their study Gibson and Yiannakis decide to sub-divide two of their 15 original tourist roles, namely *Escapist* and *Independent Mass Tourist*, into two categories each. Those

new categories are called *Escapist I* and *Escapist II* respectively *Independent Mass Tourist I* and *Independent Mass Tourist II*. In the end, 17 tourist roles are distinguished; they are listed in Table 2.1.

Name	Description
Sun Lover	*"Interested in relaxing and sunbathing in warm places with lots of sun, sand and ocean"*
Action Seeker	*"Mostly interested in partying, going to night clubs and meeting people for uncomplicated romantic experiences"*
Anthropologist	*"Mostly interested in meeting the local people, trying the food and speaking the language"*
Archaeologist	*"Primarily interested in archaeological sites and ruins; enjoys studying history of ancient civilizations"*
Organized Mass Tourist	*"Mostly interested in organized vacations, packaged tours, taking pictures/buying lots of souvenirs"*
Thrill Seeker	*"Interested in risky, exhilarating activities which provide emotional highs for the participant"*
Explorer	*"Prefers adventure travel, exploring out of the way places and enjoys challenge in getting there"*
Jet Setter	*"Vacations in elite, world class resorts, goes to exclusive night clubs, and socializes with celebrities"*
Seeker	*"Seeker of spiritual and/or personal knowledge to better understand self and meaning of life"*
Independent Mass Tourist I	*"Visits regular tourist attractions but avoids packaged vacations and organized tours"*
Independent Mass Tourist II	*"Plans own destination and hotel reservations and often plays it by ear (spontaneous)"*
High Class Tourist	*"Travels first class, stays in the best hotels, goes to shows and enjoys fine dining"*
Drifter	*"Drifts from place to place living a hippie-style existence"*
Escapist I	*"Enjoys taking it easy away from the stresses and pressures of home environment"*
Escapist II	*"Gets away from it all by escaping to peaceful, deserted or out of the way places"*
Sport Tourist	*"Primary emphasis while on vacation is to remain active engaging in favorite sports"*
Educational Tourist	*"Participates in planned study tours and seminars to acquire new skills and knowledge"*

Table 2.1: The 17 tourist roles identified by Gibson and Yiannakis and their descriptions ([GY02], p. 365).

The 17 tourist roles have been used in tourism online surveys [BDD+07]. Such survey consisted of two main parts: in the first part, participants are asked to indicate which of those roles apply to them, both at the present time and also in previous periods of their lives. The tourist roles are not explicitly mentioned but are related to statements on touristic behavioral patterns. The second part comprises 60 photos representing ten different situations related to tourism. Participants are supposed to select photos

associated with their current and past touristic behavior. In addition to that personal data and demographic characteristics of the participants are collected. The results of this study show that representative photos can be determined for almost all tourist roles. This implies that it is possible to assign a tourist role to a person based on a set of pictures that she has selected. Nevertheless, this study also shows that some tourist roles are very similar and hard to distinguish.

In [GMHF04] a study is presented that aims to relate travel personality types to travel behavioral patterns. Therefore, participants have to rate the importance of predefined travel motivations (e.g., social contact, physical activity, relaxation) and of attributes of travel destinations (e.g., scenery, good value for money, diversity). Furthermore, questions related to travel values (e.g., active vs. passive) and travel style (e.g., variety-seeking) are asked. To find out about actual travel behavior, the participants have to specify the exact destinations and activities of their most recent trip to the region under consideration (i.e., North Indiana, USA, where the study was conducted). In addition, twelve predefined travel personality types are introduced each of which are described with the help of catchy titles such as *Culture Creature* or *Beach Bum* supplemented by a short explanatory text. Those travel personality categories heve been selected as typical examples found on travel Websites. The participants have to indicate which personality type applies best to them and which least. Further data has been collected within focus groups in Chicago, Illinois. The results of the study show that the travel personalities are distinguishable regarding travel style, travel motivations and travel values. Furthermore, it turns out that if limited to the choice of one category, the majority of participants indicate the *All Arounder* as travel personality that characterizes them best. This travel type is described by the text *"You need to have it all. You go where there is a lot to do and see"* ([GMHF04], p.6). and thus, it points to a variety of travel interests. On the other hand, if this restriction is removed, i.e., more than one category can be chosen, people tend to select more than one travel type and the *All Arounder* becomes less popular. Moreover, associations between the travel personalities, on the one hand, and different preferred activities, on the other hand, are shown. However, in the study no significant relation between personality types and destination choice is detected, but here the authors argue that this might be due to specific characteristics of the region under consideration (i.e., Northern Indiana) such as its high homogeneity. The overall conclusion is, nevertheless, that travel personalities can efficiently be deployed within a destination recommender system since certain aspects such as preferences for specific activities constitute differences among travelers.

2.3.2 Social Networks

In Chapter 4 churn behavior in a Massively Multiplayer Online Role Playing Games (MMORPG) is studied with respect to social influence mechanisms. Online games provide a good opportunity to study human behavior and social interactions as they typically record players' activities on a detailed level.

A second application domain, moreover, is an online travel forum, where users discuss their upcoming trips. Goal is to investigate whether the sentiments of the users, which are obtained with the help of text-mining techniques, are inter-related.

Churn Behavior

Churn is defined as the propensity of a person to quit a team, to leave an organization, or unsubscribe a service. This term is widely applied in the context of telecommunication services. Based on the behavior of their costumers, companies have interest to differentiate between *"churners and non-churners"*; one way to assess the probability of a customer to leave the service is to assign individual *"propensity to churn"* scores ([HYW06], p. 516). However, the question of retaining users is not only relevant for the telecommunication domain, banks and insurance companies but also for Web-based services and products as their revenue is increasingly made up by revenue through advertisement. If we take Facebook as an example, adverting accounts for the main part of its revenue, and in 2015 the company gained about 17.1 billion US dollars through advertisement revenues [Sta16]. In order to maintain this important source of revenue, the regular customers have to be retained and new customers have to be attracted. For the company it is crucial to prevent churn behavior, in particular it must not happen that 80% of its users leave the platform by 2017 as a recent study claims [CS14].

In literature it has been shown that *demographic characteristics* such as gender and the indivdual socio-economic status influence the propensity to exhibit churn behavior [VdPL04]. The analysis of *user behavior* allows to draw conclusions about quitting a service or leaving a platform. Examples of this in the context of financial services are the frequency of use or early commitment [VdPL04]; in the context of the MMORPG EVE Online, the frequency of play sessions and the time between those sessions [FBS07]; in the context of the online stock brokerage E*Trade, the frequency of visiting the Websites and the number of accounts of a person [CH02].

Social interactions also influence whether or not an individual is likely to quit. The work presented in [HTK97] implies that peers and family increase customer retention through social influence or normative mechanisms. Furthermore, it has been shown that in the MMORPG World of Warcraft the social factor attracts players [DYNM06]. Social influence mechanisms are also considered in [KPS09, PUM$^+$13]. In [KPS09] a co-player network within the MMORPG EverQuest II is studied. The results imply that the probability of a player to quit depend on both her engagement in the game and the social influence from co-players. In [PUM$^+$13] quitting behavior of customers in the telecommunication domain is analyzed. It is shown that churn behavior can be predicted with higher accuracy if accumulated influence is taken into account.

Common techniques that are used to predict churn behavior are decision tree based approaches and logistic regression [DMML00, HTRR06, HYW06]. More sophisticated machine learning techniques are also applied in this context including support vector machine based approaches, random forests or neural networks [CVdP08, HYW06].

Recent work increasingly applies social network analysis techniques, such as decision tree-based classification combined with social ties [DSV+08], modified diffusion models in networks [KPS09, PUM+13] or constructing user features based on social network measures for churn prediction models [KRC+11].

Sentiments in Online Travel Forums

The role of emotions of users interacting in online forums and micro-blogging Websites is the focus of several studies. In [MPT10] Blog data is used to demonstrate that user communities emerge around certain topics. The evolution of these communities, i.e., whether they grow or shrink, is related to the emotional content of relevant posts. Posts from Blogs and BBC forums are studied in [CST+11]. This work examines how discussion evolves based on emotional contents, and it shows that the emotions of community members are likely to influence one another. In BBC online forums, where political discussions are taking place, negative emotions are dominating [CSS+11]. Connected users on the Chinese micro-blogging site Weibo show a strong sentiment correlation, especially if they interact a lot. However, negative emotions seem to have a higher impact than positive emotions [FZCX14]. Instead, in the context of MySpace comments positive emotions appear to have a higher impact [TWU10]. It was also observed that there are clear gender differences. Female users express positive emotions more often than male users. In [KGH14], the so-called "Facebook Study", experimental evidence for massive-scale contagion of emotional content on Facebook is given. In the study, the messages that are displayed to the users are filtered in a way that some users receive less positive contents and some less negative. It turns out that the users start to behave accordingly in their own messages, i.e., they produce fewer positive and negative contents accordingly.

In order to assess the sentiment of user-generated content, supervised machine learning methods are commonly used. However, [KGH14] applies a lexical-based approach, which is also done in our work. To study correlations and interdependencies between user sentiments various techniques are used, such as temporal approaches including time series and diffusion models [MPT10, FZCX14], agent based models [CSS+11], anova tests [TWU10], conditional probabilities [CST+11], and regression methods [KGH14].

In the context of tourism often lexical-based sentiment analysis is chosen to quantify the emotionality of a text or a user. The term sentiment analysis refers to approaches that aim to extract subjectivity from text either to decide whether a text is objective or subjective, or whether a subjective text is positive or negative. The lexicon-based approach utilizes sentiment dictionaries to quantify the subjective of a text by aggregating the sentiments assigned to the words in that text [TBT+11]. In [GZFF12] a lexicon-based approach is applied to relate tourism related reviews to their numerical rating. Using such an approach, the authors are able to classify reviews as *"good"* or *"bad"* in a quite accurate way. In [SHFL13] statements about product properties of hotel reviews are extracted. The statements are tested to determine if they are subjective, and if so, whether they are positive or negative. The authors show that for subjectivity recognition the lexical

based approach performs better than various supervised machine learning techniques. In [GSO11] an approach is introduced that makes use of lexical data bases to calculate sentiment scores of tourism related reviews.

2.3.3 Multi-Level Systems

To examine, how all three levels of information, i.e., the individual level, the group level and the network level, might be combined within one model, we study team performance in complex head-to-head competitions. The following review of team literature is based on [NHC15].

Team Performance

Most contemporary challenging tasks need to be addressed by teams. As a consequence, there is an increasing interest in studying how teams form and how that affects their performance. There is a small but growing body of work that aims at identifying team assembly factors that affect team performance in contexts as diverse as scientific collaborations and Broadway musicals [GUSA05].

A team can be described as a complex collection of individuals and their interactions. Therefore, many factors influence team processes and their outcomes. In [GD96] the research on groups and teams in organizations is reviewed, they also exmine factors that influence their effectiveness such as group composition, cohesiveness, leadership, and motivation. Furthermore, in [FBR99] a set of social-psychological factors for effective virtual teams is proposed. By extending Hackman's Model of Group Effectiveness to teamwork in virtual environments these factors are classified into the general categories: organizational context, group design, group synergy, group process, and group material resources.

Based on the literature, we focus on three categories of factors in our team performance analysis: compositional, relational, and team ecosystem. The first two categories, i.e., compositional and relational factors, are clearly defined and well characterized in literature. The research on inter-team relations, on the other hand, is limited. To address relations between teams explicitly and to capture their impact on team performance in more detail, we propose team ecosystem factors as a separate category.

Considering a team as a collection of individuals, *compositional factors* measure team members' attributes such as their personal characteristics and their capabilities and knowledge related to team activities. The literature on virtual teams and team performance covers compositional factors in depth; many studies focus especially on task-related individual attributes such as skills or expertise of the team members to illustrate how compositional aspects influence the effectiveness of a team [CB97, CK87]. The impact of task-related compositional factors is straightforward: teams with higher skilled members are more task-cohesive, and therefore, more likely to succeed than the teams with less competent members.

The importance of *relational aspects* within teams has been studied for several decades. In [BH06] a meta-analysis based on 37 studies is conducted to find out whether and how network structure impacts team effectiveness (i.e., team viability and performance). The examinations of hypotheses related to density-performance, density-viability, match of tie content to team outcomes, centrality-performance, and moderating effects of time shows that teams with denser network structures tend to perform better. This is true for both instrumental ties (i.e., ties that emerge from formal relationships) and expressive ties (e.g., ties that reflect friendship); and both types of ties have similar predictive power. Furthermore, teams with denser networks tend to have greater team viability. Again, this is true for both instrumental and expressive ties. However, here expressive ties are a stronger predictor for team viability than instrumental ties. Further analysis considers two distinct aspects of the moderating effects of time. First, it is shown that there is a causal sequencing of network structure and team performance; network structure is antecedent to team performance rather than vice versa. Second, the impact of network structure on performance declines with time; the more the team members get familiar with each other and their tasks, the weaker is the effect of ties on performance. Thus, overall the meta-analysis shows that network relations clearly impact team performance and team viability.

In [MC00] global virtual task teams, their dynamics and their effectiveness are observed in a qualitative study over a period of 21 months. The study reveals that the effectiveness of a global virtual team is related to a series of adequate communication incidents among team members. In [PNCM+13] and [PNCMW13] the positive effects of players' previous teaming relations and friendship on team's winning chances in the Multiplayer Online Battle Arena (MOBA) game Dota 2 is illustrated.

Among social relations, team collaboration history, i.e. previous collaboration relations, are particularly important to explain team performance [JR09]. When forming new teams, people often prefer partners they are familiar with from previous work or joint projects. Furthermore, as explained by the *"performance-outcome learning"* perspective [SM08], previous performance outcomes also influence the chances of future collaboration. Individuals with successful previous collaboration are more likely to team up again in future activities. Prior knowledge reduces uncertainty; prior success increases the social capital of the team members, which in turn enhances the outcome of new collaboration [AU07].

Teams are not standalone entities in an organization [AC92]. Teams learn from both internal and external sources and are influenced by team members' external experiences and relations [ABC09]. When members of different teams work together, their collaborations establish relationships between these teams, which in turn results in a complex *team ecosystem*. Within-team and between-team relationships and their impact on the performance of teams of students are studied in [dMSS+14]. Two types of relations are considered: expressive ties (represented by friendship relations) and instrumental ties (defined by the time that collaborators spend in physical proximity). Results show that only the strong ties (expressive as well as instrumental ties) have an impact on the performance of the team and the impact is significant for both within-team and

between-team relations. The positive effects of inter-team relations on team performance has also been observed in another setting with teams of students [BBJ97], R& D project teams [WHG01], and work groups in organizations [OCL04].

In a team ecosystem, membership overlaps between teams (*"structural folding"*) might significantly contribute to the higher performance of a team and its creative success [dVSV14]. Similarly, in [BM14] the concept *"network oscillation"*, which refers to an iterative process of deep engagement in a group and brokering across groups, is introduced to characterize the network advantage on performance. Inter-team connections form social capital [Bur00] and bring in new knowledge for the benefit of the team [RWE+09].

Team-vs-Team Setting

In many of the settings previously described, the detailed process of team collaboration depends on the nature of the tasks. In team-vs-team competitions, on the other hand, there are typically no pre-defined tasks; the overall objective is rather to defeat the opponent. Thus, in these settings, the winner is clearly identified but is based on relative performance vis-a-vis the loser. Furthermore, a team has to react constantly to the opponent's activities. This helps to reflect team internal dynamics from a more general perspective and provides the opportunity to study team performance in a comprehensive way.

Team-vs-team competitions often occur in sports but also in other areas such as business (e.g., the competition of two standards) and of course in military conflicts. However, by now these settings have drawn little attention from researchers, there are only very few articles explicitly modeling team-vs-team competitions. In [Kle92], for instance, head-to-head competitions between companies with similar product lines are studied. With the help of a mathematical model it is shown that matching product lines might lead to a less competitive market since using various suppliers is more costly for customers (e.g., extra effort) which they try to avoid. Another application is presented in [DWA10], where the performance of soccer players in the Euro Cup 2008 tournament is examined and a network approach is applied to quantify both the performance of a team as well as the contributions of individual players. Furthermore, it is illustrated how the introduced method could be generalized to other settings such as scientific collaborations. In [MC14] scoring dynamics in professional team-sports competitions are analyzed. A generative model that only takes into account tempo and balance of scoring events is developed to accurately predict the outcome of a match. Our research seeks to build on this tradition.

As a digital replica of real world scenarios, virtual communities and online games may provide an opportune environment to study human behavior and interactions [Wil10, WCP+11]. Research in this context often studies players interacting with each other and collaborating in game teams in MMORPGs. Most of the studies focus on in-game organizations or groups aiming to finish specific tasks, called Player versus Environment (PvE). For example, in [WDX+06] social communities in World of Warcraft (so-called guilds) are analyzed and it is shown that due to the game's design some interactions are

encouraged whereas others are discouraged. Player versus Player (PvP) play style games show the advantage to directly measure the outcome of two teams competing with each other. Typically PvP games have no predetermined tasks and some-times the game rules are flexible and changeable according to players' strategies. In [CH12] dynamics related to the development of game rules and the emergence of new variants are examined for the two popular online PvP games Texas Hold'em and Halo 2.

In recent years, MOBA games in which two teams combat in a standard battlefield have become very popular and further developed into a type of electronic sports (e-sports). Similar to other team sports such as basketball and soccer, more and more professional competitions are taking place and winning becomes increasingly lucrative. As a consequence, research has begun to focus on the competitive characteristic of online games. In [CG13] EVE online is studied. This game encourages ruthless play styles and unsocial behavior and it is discussed how the e-sports version of the game could account for these unique features. In [WS13] competitive virtual environments and customer needs are analyzed. E-sports services are bringing cooperation and competition together. In the study a connection is established between competitive need gratification and hedonic need gratification, on the one hand, and e-sports use, on the other hand. Good predictors for the use of e-sports are competition and challenge (related to competitive need gratification) and escapism (related to hedonic need gratification). However, social relationship and fun (also related to hedonic need gratification) do not have a significant effect in this study. Playing style and performance in the MOBA game Dota 2 are used in [GJWL13] to classify the role of a player within the team. Here, also the increasing popularity of e-sports events is discussed.

CHAPTER 3

From Individuals to Collective Preferences

The goals of this chapter are to compare the individual level and the group level (see part A and B in Figure 3.1) and to study social influence processes and the impact of social context on human behavior with respect to these levels. Applying the GDA approach, we discuss the construction of a social space based on collective preferences. Insights into the social system as a whole and into individual opinions and behaviors are gained by interpreting the dimensions of this space and by studying the positions of the individuals. This approach emphasizes common characteristics among the individuals. However social interactions are not explicitly taken into account. The social influence processes, which are addressed, are comparison mechanisms.

In this chapter, we will first use a data set from literature on teenage smoking behavior to illustrate the individual level and the construction of the group level. Then we will use this approach in the context of picture-based recommender systems. It will enable us to construct a space based on individual preferences that comprises both the user profiles and products. Finally, we show with the help of statistical analyses that this representation captures the setting in an accurate way.

3.1 Metric Spaces for Individual Behavior

Now we discuss models on the individual and the group level. To compare those two levels and illustrate their differences, we use a data set on the smoking behavior among teenagers from the literature. The main goal of the *Teenage Friends and Lifestyle Study* [MW96, Mic97, PW03] was to examine the smoking behavior of teenagers and to identify processes that influence their attitudes towards smoking. From the beginning of 1995 to 1997, data of school children in Glasgow were collected at three points in time t_1,

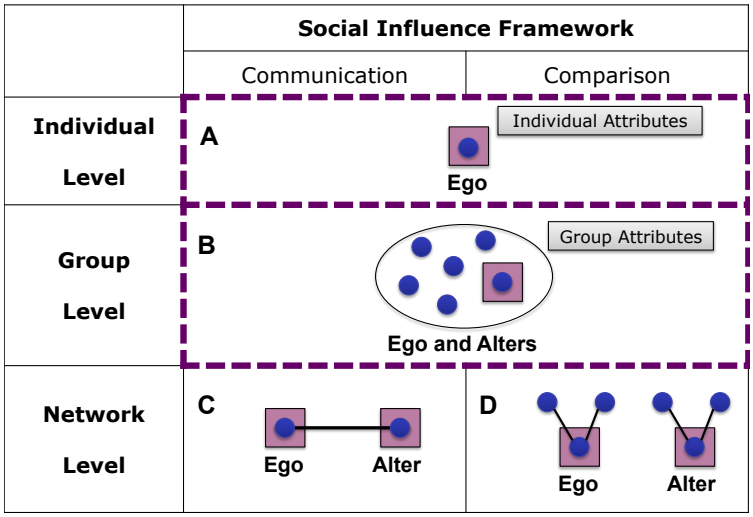

Figure 3.1: Social influence processes and levels of information.

t_2 and t_3, always at an interval of one year. Overall, 160 pupils were interviewed about demographic characteristics, their smoking, drinking and cannabis usage behavior as well as lifestyle preferences. At the beginning of the study, the pupils were about 13 years old. The data is online available [Tom15a].

3.1.1 Glasgow Data Set

Since our models are cross-sectional, we base our discussion on the behavior and preferences indicated at only one time point. Here we choose the second time point t_2, i.e., January 1996, because later in the discussion we will use information from the past, i.e., time point t_1, and predict future behavior, i.e., the behavior at t_3.

Relevant Individual Attributes

Out of the 160 pupils, 76 (47.5%) are female and 84 (52.5%) are male. In January 1996 the pupils are between 13.4 and 15.6 years old. The average age is 14.35 years with a standard deviation of 0.33. The age of one pupil is unknown.

Information on smoking, drinking and cannabis usage behavior is missing for some individuals at time point t_2 (for 14 pupils it is unknown whether they consume tobacco or cannabis at t_2 and for 36 pupils it is unknown whether they drink alcohol at that time). For these variables we conduct data imputation. The reason is that later in the discussion when we consider the friendship network among the pupils this information helps to gain a more comprehensive view of the setting.

Regarding their smoking behavior, the pupils can indicate if they do *not smoke*, if the smoke *occasionally* or if they smoke *regularly*, i.e., more than once a week. Drinking is measured by *no drinking*, *once or twice a year*, *once a month*, *once a week* and *more*

than once a week, and cannabis consumption by *no cannabis*, *tried once*, *occasionally* and *regularly*. In the cases of missing data, we consider the behavior of the individual at the other time points. For example, if the smoking behavior of a pupil is missing at t_2 and also t_1, but this pupil indicates that she is smoking regularly at t_3 , we also assume this behavior at t_2. If only the behavior at t_2 is missing but the behaviors at t_1 and t_3 are known and different, we assume the "stronger", i.e., the more frequent, behavior at t_2. For example, if a pupil, whose behavior at t_2 is unknown, indicates that she does not smoke at t_1 but that she smokes occasionally at t_3 , we also assume that she smokes occasionally at t_2. For all pupils information on smoking and cannabis consumption is available for at least one time point. For six pupils, on the other hand, the drinking behavior is missing at all three points in time. Here we assign the middle category, i.e., once a month.

	Yes	No
Female	76 (47.5%)	84 (52.5%)
Smoking	37 (23.1%)	123 (76.9%)
Drinking	46 (28.8%)	114 (71.2%)
Cannabis	29 (18.1%)	131 (81.9%)
Relation	41 (25.6%)	119 (74.4%)
Sibling Smokes	27 (16.9%)	133 (83.1%)
$N = 160$		

Table 3.1: Distribution of binary individual attributes.

After dealing with missing data as described, the distributions of the pupils' substance usage behaviors are as follows: Regarding smoking tobacco, 123 pupils (76.9%) indicated that they do not smoke, 10 (6.2%) that they smoke occasionally and 27 (16.9%) that they smoke regularly (i.e., more than once a week). Regarding alcohol, 9 pupils (5.6%) do not drink at all, 56 (35.0%) drink once or twice a year, 49 (30.6%) drink once a month, 32 (20.0%) drink once a week and 14 pupils (8.8%) drink alcohol more than once a week. Regarding cannabis, the majority (101 pupils or 63.1%) has never tried it and 30 pupils (18.8%) answered that they tried it once. On the other hand, 26 pupils (16.2%) occasionally use cannabis and 3 pupils (1.9%) use it regularly.

Furthermore, since later in the discussion when looking at social influence network models, the outcome variable is required to be binary, we dichotomize the behavior that we aim to study. For smoking tobacco, we merge the categories *occasionally* and *regularly* to indicate whether a person is a smoker; and a pupil is a non-smoker if she indicated it. Although those variables will not serve as outcome variables, we also dichotomize alcohol and cannabis consumption to make them better comparable to smoking behavior. Regarding drinking, we merge the categories *once a week* and *more than once a week* to indicate that a person exhibits drinking behavior. On the other hand we merge *no drinking*, *once or twice a year* and *once a month* to indicate that drinking behavior is not present. Regarding cannabis, we merge the categories *occasionally* and *regularly* to indicate that a person consumes cannabis. On the other hand, we merge *no cannabis*

and *tried once* to capture that a person does not consume it. The resulting distributions of substance usage behaviors are: 37 pupils (23.1%) smoke tobacco compared to 123 pupils (76.9%) who do not smoke tobacco; 46 pupils (28.8%) drink compared to 114 pupils (71.2%) who do not drink; and 29 pupils (18.1%) consume cannabis compared to 131 pupils (81.9%) who do not consume cannabis.

The pupils were also asked if they were in a romantic relation. At time point t_2, this information is missing for 14 individuals. For those individuals we assume the same status as at time point t_1 and/or t_3. Here, no ambiguities occur. Out of the 160 pupils, 41 (25.6%) are in a romantic relationship and 119 (74.4%) are not. We also consider whether a pupil has at least one smoking sibling. This information is missing in 10 cases. In these cases we consider the smoking behavior of the parents, since the information for different time points is not provided. This results in 27 pupils (16.9%) who have at least one smoking sibling compared to 133 (83.1%) have no smoking sibling.

The distributions of the binary user attributes after data imputation are summarized in Table 3.1. In Table 3.2 the correlations between those attributes are listed; here also the age of the pupils is included, where we substituted the one missing value by the average age of all pupils. Being female is weakly correlated with smoking behavior (0.16, $p < 0.05$) and moderately correlated with being in a romantic relation (0.24, $p < 0.01$). The age of a pupil is only correlated with consuming cannabis (0.21, $p < 0.01$). Smoking behavior is correlated with all other attributes but age. It is moderately correlated with having a smoking sibling (0.25, $p < 0.01$) and being in a romantic relation (0.26, $p < 0.01$). Also with drinking behavior it is moderately correlated, but here the correlation is strongly significant (0.34, $p < 0.001$). The highest correlation coefficient can be found between smoking and the consumption of cannabis (0.53, $p < 0.001$). Also drinking behavior is moderately correlated with cannabis consumption (0.40, $p < 0.001$) and weakly correlated with both being in a romantic relation (0.18, $p < 0.05$) and having a smoking sibling (0.18, $p < 0.05$). These two attributes are also correlated with the consumption of cannabis (0.25, $p < 0.01$ and 0.28, $p < 0.001$ respectively) and with each other (0.16, $p < 0.05$).

Further attributes on demographic characteristics, pocket money and family smoking behavior that are contained in the data set are not taken into account, since they have turned out not to be relevant for our analysis.

Leisure Activities

The pupils were also asked about their leisure time. They could indicate how often they participate in various pre-defined activities, namely, *"I spend time on my hobby (eg art, an instrument)"* (abbreviated as *art*), *"I do nothing much (am bored)"* (abbreviated as *bored*), *"I go to church, mosque or temple"* (abbreviated as *church*), *"I go to cinema"* (abbreviated as *cinema*), *"I go to dance clubs or raves"* (abbreviated as *clubs*), *"I go to pop concerts, gigs"* (abbreviated as *concerts*), *"I go to sport matches"* (abbreviated as *matches*), *"I listen to tapes or CDs"* (abbreviated as *music*), *"I play computer games"* (abbreviated as *PC games*), *"I look after a pet animal"* (abbreviated as *pet*), *"I read*

	Female	Age	Smoking	Drinking	Cannabis	Relation	Sibling Smokes
Female	1.00						
Age	0.02	1.00					
Smoking	0.16*	0.12	1.00				
Drinking	0.10	0.05	0.34***	1.00			
Cannabis	0.00	0.21**	0.53***	0.40***	1.00		
Relation	0.24**	0.08	0.26**	0.18*	0.25**	1.00	
Sibling Smokes	0.11	0.08	0.25**	0.18*	0.28***	0.16*	1.00

Note: *p<0.05; **p<0.01; ***p<0.001; $N = 160$.

Table 3.2: Correlation table of individual attributes.

comics, mags or books" (abbreviated as *read*), *"I go to something like B.B., Guides or Scouts"* (abbreviated as *scouts*), *"I look around in the shops"* (abbreviated as *shops*), *"I take part in sports"* (abbreviated as *sports*), and *"I hang round in the streets"* (abbreviated as *streets*). At time t_2, for eight of the variables 14 values are missing and for seven of the variables 15 values are missing. Here we apply the same strategy for data imputation as described previously, i.e., we are using the information from time t_1 and/or t_2. If ambiguities occur, we select the more frequent behavior.

In Table 3.3 the distributions of the answers are displayed. 82 pupils (51.2%) spend time on arts or an instrument once a month or less. On the other hand, 78 pupils (48.8%) pursue this activity at least once a week. The vast majority of pupils is rarely bored (130 or 81.3%) and very seldom visits a church, mosque or temple (134 or 83.8%). Typically the pupils go to the movies once a month (103 or 64.4%). However, 34 pupils (21.3%) indicate that they do it at least once a week. The majority of pupils visits clubs or raves (101 or 63.1%) and concerts (126 or 78.8%) less than once a month. Going to sport matches is quite popular for 48 pupils (30.0%), they do it at least once a week. On the other hand, 75 pupils (46.9%) visit sport matches less than once a month. The most popular activity is listening to music; 142 pupils (88.8%) indicate that they do it almost every day. Playing PC games is a quite popular activity for 90 pupils (56.3%), the do it at least once a week. On the other hand, 55 pupils (34.4%) play computer games less than once a month. The answers regarding looking after a pet are quite dichotomous, pupils either do it almost every day (78 or 48.8%) or less than once a month (71 or 44.4%); only 11 pupils (6.9%) indicate something else. Reading is very popular; 116 pupils (72.5%) do it at least once a week. However, 44 pupils (27.5%) indicate that they read once a month or less. Going to Scouts is not very popular for most of the pupils; 114 (71.2%) do it less than once a month. Pupils typically go shopping once a week (110 or 68.8%). Half of the pupils (80 or 50.0%) do sports most days and 42 pupils (26.2%) at least once a week. However, 23 pupils (14.4%) indicate that they participate in sports less than once a month. Another quite popular activity is hanging around in the streets; 88 pupils (55.0%) do it most days. On the other hand, 43 pupils (26.9%) indicate that they do it less than once a month.

	Less than Once a Month	Once a Month	Once a Week	Most Days
Art	66 (41.2%)	16 (10.0%)	43 (26.9%)	35 (21.9%)
Bored	130 (81.2%)	8 (5.0%)	12 (7.5%)	10 (6.2%)
Church	134 (83.8%)	6 (3.8%)	19 (11.9%)	1 (0.6%)
Cinema	23 (14.4%)	103 (64.4%)	28 (17.5%)	6 (3.8%)
Clubs	101 (63.1%)	39 (24.4%)	13 (8.1%)	7 (4.4%)
Concerts	126 (78.8%)	29 (18.1%)	3 (1.9%)	2 (1.2%)
Matches	75 (46.9%)	37 (23.1%)	41 (25.6%)	7 (4.4%)
Music	7 (4.4%)	1 (0.6%)	10 (6.2%)	142 (88.8%)
PC Games	55 (34.4%)	15 (9.4%)	39 (24.4%)	51 (31.9%)
Pet	71 (44.4%)	3 (1.9%)	8 (5.0%)	78 (48.8%)
Read	23 (14.4%)	21 (13.1%)	49 (30.6%)	67 (41.9%)
Scouts	114 (71.2%)	1 (0.6%)	32 (20.0%)	13 (8.1%)
Shops	11 (6.9%)	19 (11.9%)	110 (68.8%)	20 (12.5%)
Sports	23 (14.4%)	15 (9.4%)	42 (26.2%)	80 (50.0%)
Streets	43 (26.9%)	12 (7.5%)	17 (10.6%)	88 (55.0%)

$N = 160$

Table 3.3: Frequencies of pupils' leisure time activities.

The correlations between the leisure activities are displayed in Table 3.4. Overall, most leisure activities are not correlated. However, there are some significant weak and moderate correlations. Spending time on arts or an instrument is weakly correlated with reading (0.16, $p < 0.05$) and moderately correlated with going to Scouts (0.21, $p < 0.01$) and going to church, mosque or temple (0.22, $p < 0.01$). Furthermore, there is a moderate negative correlations between dedicating time to arts or an instrument and hanging around in the streets (-0.33, $p < 0.001$). Being bored is nor correlated to any other leisure time activity, which is reasonable. Going to church, mosque or temple in the free time is positively correlated with going to Scouts (0.32, $p < 0.001$) and negatively correlated with hanging around in the streets (-0.23, $p < 0.01$). The activity going to the cinema is positively correlated with a number of other activities, namely visiting sport matches (0.20, $p < 0.05$), shopping (0.22, $p < 0.01$), playing computer games (0.26, $p < 0.001$), visiting clubs or raves (0.32, $p < 0.001$) and going to concerts (0.33, $p < 0.001$). Visiting clubs or raves is weakly correlated with hanging around in the streets and strongly correlated with going to concerts (0.47, $p < 0.001$). The latter also represents the highest correlation between the leisure time activities. However, visiting clubs or raves is weakly negatively correlated with participating in sports (-0.19, $p < 0.05$) and going to Scouts (-0.20, $p < 0.05$). Also going to concerts is negatively correlated with sports (-0.16, $p < 0.05$). Visiting sport matches is moderately correlated with playing computer games (0.23, $p < 0.01$) and with actively participate in sports (0.34, $p < 0.001$). On the other hand, it is clearly negatively correlated with reading (-0.33, $p < 0.001$). Listening to music is weakly correlated with hanging around in the streets (0.17, $p < 0.05$). Playing computer games, on the other hand, is moderately negatively correlated with hanging around on the streets (-0.22, $p < 0.01$). However, it is

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1 Art	1.00														
2 Bored	-0.13	1.00													
3 Church	0.22**	-0.08	1.00												
4 Cinema	-0.11	-0.08	-0.03	1.00											
5 Clubs	-0.01	-0.09	-0.09	0.32***	1.00										
6 Concerts	-0.08	-0.13	-0.05	0.33***	0.47***	1.00									
7 Matches	-0.06	-0.15	-0.09	0.20*	0.10	0.13	1.00								
8 Music	-0.09	0.01	-0.14	0.10	0.04	0.08	0.00	1.00							
9 PC Games	0.07	-0.09	0.00	0.26***	-0.13	-0.05	0.23**	0.00	1.00						
10 Pet	0.13	0.10	-0.01	-0.02	-0.03	0.02	-0.13	0.01	0.09	1.00					
11 Read	0.16*	-0.04	0.04	-0.12	-0.09	0.00	-0.33***	0.14	0.01	-0.01	1.00				
12 Scouts	0.21**	-0.11	0.32***	-0.08	-0.20*	-0.10	0.05	-0.11	0.22**	0.10	0.11	1.00			
13 Shops	0.00	-0.11	0.13	0.22**	0.07	0.15	-0.13	0.12	0.02	0.09	0.02	-0.06	1.00		
14 Sports	0.06	-0.10	0.10	0.02	-0.19*	-0.16*	0.34***	0.05	0.21**	0.00	-0.12	0.18*	-0.04	1.00	
15 Streets	-0.33***	0.05	-0.23**	0.01	0.18*	0.08	0.04	0.17*	-0.22**	-0.17*	-0.06	-0.16*	0.24**	-0.06	1.00

Note: *p<0.05; **p<0.01; ***p<0.001; $N = 160$.

Table 3.4: Correlation table of leisure activities.

positively correlated with participating in sports (0.21, $p < 0.01$) and going to Scouts (0.22, $p < 0.01$). Looking after a pet is negatively correlated with hanging around in the streets (-0.17, $p < 0.05$). Furthermore, going to scouts is weakly positively correlated with sports (0.18, $p < 0.05$) and weakly negatively correlated with hanging around in the streets (-0.16, $p < 0.05$). Finally, shopping is moderately positively correlated with hanging around in the streets (0.24, $p < 0.01$).

Furthermore, we take a look at the correlations between smoking and the pupils' leisure activities. Here it turns out that there is a moderate positive correlation between smoking and hanging around in the streets (0.25, $p < 0.01$) and going to clubs (0.32, $p < 0.001$) respectively. On the other hand, smoking is negatively correlated with sports (-0.18, $p < 0.05$) and playing computer games (-0.21, $p < 0.01$). There is no other significant correlation between smoking and the discussed leisure activities.

3.1.2 Individual Attributes and Smoking

If we want to understand why a person is smoking, it is an obvious approach to take the individual attributes of this person into account. To model the individual smoking behavior based on the attributes of the individual, we use binary logistic regression (see Section 2.2.1). As discussed in Section 3.1.1 some attributes can clearly be associated with smoking behavior (see also Table 3.2). Based on these insights, we develop different models. In some of them also those leisure activities that have a correlation with smoking (as discussed at the end of Section 3.1.1) are considered. In all models, the binary smoking behavior serves as outcome variable. In Table 3.5 the results are shown.

In Model 1 we select as predictor variables those attributes from Table 3.2 that are correlated with smoking but do not capture the pupil's substance usage behavior or leisure time activities, i.e., we include the gender, whether a pupil is in a romantic relationship and whether she has a smoking sibling. In Model 2 we also include a pupil's drinking behavior as predictor variable and in Model 3 additionally whether a pupil consumes cannabis. In Model 4 the leisure activities playing computer games, being a scout, doing sports and hanging around in the streets are included instead of drinking alcohol and cannabis. Finally, in Model 5 all the mentioned predictor variables are taken into consideration. To better see the different effects that might exist, we choose a p-value of 0.1 for the discussion of the results.

The binary predictor being female leads to a 225% odds ratio to smoke in the first model ($p < 0.1$), to a 205% odds ratio to smoke in the second model ($p < 0.1$) and to a 263% odds ratio to smoke in the third model ($p < 0.05$). In the other models, which also take into consideration the leisure activities, the effect of this predictor completely disappears. Being in a romantic relation shows a significant positive impact on smoking behavior in three of the models. Having a boy- or a girlfriend leads to a 258% odds ratio to smoke in Model 1 ($p < 0.05$), to a 233% odds ratio to smoke in Model 2 ($p < 0.1$) and to a 265% odds ratio to smoke in Model 4 ($p < 0.05$). However, the effect disappears when the consumption of cannabis is included as a predictor. Having a smoking sibling shows

a significant positive impact on smoking behavior in the first model; there it leads to a 282% odds ratio to smoke ($p < 0.05$). In the second model it is marginally significant, here it leads to a 244% odds ratio to smoke ($p < 0.1$). In the other models, no effect of smoking siblings is detected.

Drinking and consuming cannabis have a strong impact on smoking behavior. Drinking leads to a 413% odds ratio to smoke in Model 2 ($p < 0.01$). This effect decreases if cannabis is included as predictor variable. Then, drinking alcohol leads to a 315% odds ratio to smoke in Model 3 ($p < 0.05$) and to a 298% odds ratio to smoke in Model 5 ($p < 0.05$). Using cannabis is the strongest predictor. It leads to a 960% odds ratio to smoke in Model 3 ($p < 0.01$) and to a 817% odds ratio to smoke in Model 5 ($p < 0.01$).

Regarding the leisure activities only hanging around in the streets shows a significant effect in Model 4; doing this more frequently has a positive impact on the likelihood of smoking. One higher frequency category leads to a 168% odds ratio to smoke ($p < 0.05$). However, this effect disappears when drinking and using cannabis are included (see Model 5 in Table 3.5). The constant term is strongly significant in all models but Model 4.

All listed criteria to assess the quality of the models, i.e., Pseudo R^2, Log Likelihood and Akaike's Information Criterion, show that apart from Model 5, Model 3 captures the setting better than the other models. Model 5 contains all predictor variables and thus obviously fits the data best. However, there is no big gain in quality compared to Model 3 but many more variables included. For example, the Pseudo R^2 in Model 1 is 16.8%. This increases to 25.9% when drinking as predictor variable is added (see Model 2). However, in Model 3 cannabis is included additionally. Here, the Pseudo R^2 increases to 39.6%. Although more predictor variables are included in Model 4, it accounts for less variance (the Pseudo R^2 is 28.4%). It is obvious that Model 5 has the highest Pseudo R^2, namely 42.5%, as it contains all predictors of the other four models.

We see that the behavior of a person can be captured by attributes of this person that are related to demographic characteristics, other behaviors and activities, their personal status, and so on. However, here no relations between the individuals or groups of individuals are taken into account. All pupils in the study are considered to be independent. However, one way to go beyond this individual level is to explicitly focus on collective behavior and preferences.

3.1.3 Collective Behavior and Smoking

To take social context into consideration in order to understand how the pupils might influence one another, we choose a geometric data analysis approach (see Section 2.2.2). With this approach we will analyze whether certain individual attributes can explain the position of a pupil in the social space as well as whether a position indicates a certain behavior.

To construct the metric space to capture social context, we conduct a factor analysis based on the pupils' leisure activities. The idea of this approach is to identify latent

	Dependent variable:				
	Smoking				
	(1)	(2)	(3)	(4)	(5)
Female	0.811*	0.719*	0.966**	0.088	0.777
	(0.419)	(0.436)	(0.489)	(0.504)	(0.584)
Relation	0.949**	0.847*	0.741	0.973**	0.711
	(0.428)	(0.449)	(0.499)	(0.467)	(0.522)
Sibling Smokes	1.036**	0.892*	0.357	0.826	0.299
	(0.475)	(0.502)	(0.584)	(0.509)	(0.590)
Drinking		1.418***	1.146**		1.093**
		(0.423)	(0.466)		(0.490)
Cannabis			2.262***		2.100***
			(0.547)		(0.621)
PC Games				-0.262	0.147
				(0.187)	(0.226)
Scouts				-0.287	-0.231
				(0.249)	(0.274)
Sports				-0.206	-0.148
				(0.199)	(0.229)
Streets				0.521**	0.379
				(0.207)	(0.231)
Constant	-2.157***	-2.568***	-3.000***	-1.733	-3.598***
	(0.347)	(0.396)	(0.458)	(1.154)	(1.363)
Pseudo R^2	16.8%	25.9%	39.6%	28.4%	42.5%
Log Likelihood	-77.116	-71.495	-62.281	-69.865	-60.137
Akaike Inf. Crit.	162.233	152.991	136.562	155.730	140.274
N	160	160	160	160	160

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3.5: Logistic regression models.

factors that represent underlying relationships between the considered variables (see also Section 2.2.2). The correlations between the leisure activities are displayed in Table 3.4. As discussed previously, most of the variables are hardly correlated. However, there are some modest correlations.

We apply the Kaiser-Meyer-Olkin (KMO) criterion to assess whether the data is suitable for factor analysis [Bü10]. Based on this criterion we remove the activities looking after a pet and shopping from our analysis as they have both a very low measure of sampling adequacy (MSA). After that each variable has a MSA of at least 0.5. Furthermore, the overall MSA of 0.61 can be regarded as acceptable. In the scree plot in Figure 3.2, it is shown that five eigenvalues are larger than one. Based on this, we extract five factors using maximum likelihood with varimax rotation (see also Section 2.2.2).

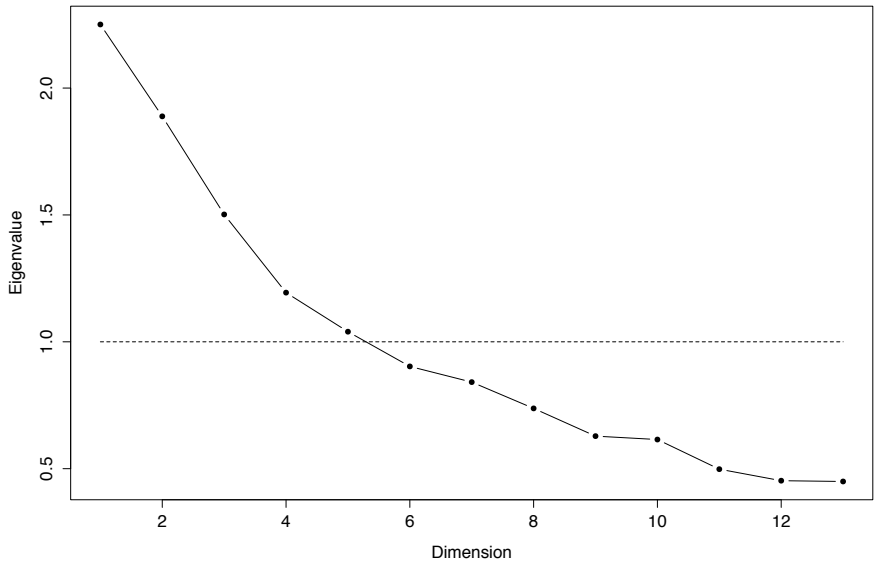

Figure 3.2: Leisure activities: scree plot.

In Table 3.6 the loadings of the factors are displayed (we only show loadings with an absolute value greater than 0.2 to underline the effects more clearly). The first factor is mainly positively determined by the activity going to clubs or raves (loading 0.69), going to concerts (loading 0.68), going to the cinema (loading 0.49). Also attending sport matches plays a role (loading 0.21). Being bored on the other hand is negatively related with this factor (loading -0.20). We summarize the factor, that accounts for 10% of the variance, as *Going-Out*. Factor 2 is strongly defined by going to the church, mosque or temple (loading 0.53), spending time on arts or an instrument (loading 0.51), and going to Scouts (0.44). On the other hand, there are also two variables that are negatively related with this factor, namely hanging around in the streets (loading -0.49) and listening to music (loading -0.27). This factor accounts for 9% of the variance and we call it *Arts &*

Religion. Factor 3 is mainly defined by playing PC Games (loading 0.97) and going to the cinema (loading 0.26), and it accounts for 8% of the variance. We call it *Games & Movies*. Factor 4 is defined by participating in sports (loading 0.67) as well as attending sport matches (loading 0.54). It accounts for 7% of the variance and we call it *Sports*. The fifth factor is mainly determined by reading (loading 0.89) and listening to music (loading 0.23). Attending sport matches, on the other hand, is negatively related to this factor (loading -0.22). It accounts for 7% of the variance and we call it *Reading & Music*.

Thus, the five factors account for 42% of the variance in the data. In general, this can be seen as rather modest. However, here higher factor solutions, where more variance is explained, lead to redundant, i.e., very similar factors. The interpretability suffers in these cases. Thus, for our purposes it clearly makes sense to base the further discussion on the five factor solution. In particular, the hypothesis that five factors are sufficient cannot be rejected (the p-value is 0.719). Thus, the model fits the data adequately.

	Going-Out	Arts & Religion	Games & Movies	Sports	Reading & Music
Art		0.51			
Bored	-0.20				
Church		0.53			
Cinema	0.49		0.26		
Clubs	0.69				
Concerts	0.68				
Matches	0.21			0.54	-0.22
Music		-0.27			0.23
PC Games			0.97		
Read					0.89
Scouts		0.44			
Sports				0.67	
Streets		-0.49			

Table 3.6: Factor analysis: loadings of the five factor solution (only the loadings greater than 0.20 or smaller than -0.20 are displayed).

We can also visualize the leisure activities in the constructed space. In Figure 3.3 all 13 activities that have been considered in the factor analysis are located with respect to the dimensions *Going-out* and *Arts & Religion*. Activities that are located along a dimension strongly characterize this dimension. Furthermore, activities that are close are similar with respect two the two dimensions. Such pictures help to find accurate interpretations for the setting.

In Figure 3.4 the described relations between the activities and the factors are visualized. Positive loadings are shown in blue, negative loadings are shown in red. The higher the loading the thicker and darker the connection line. The labels are three characters abbreviations of the variable names (i.e., *"Bored"* is abbreviated as *"Brd"*, *"Church"* as *"Chr"*, *"Cinema"* as *"Cnm"*, *"Clubs"* as *"Clb"*, *"Concerts"* as *"Cnc"*, *"Matches"* as *"Mtc"*,

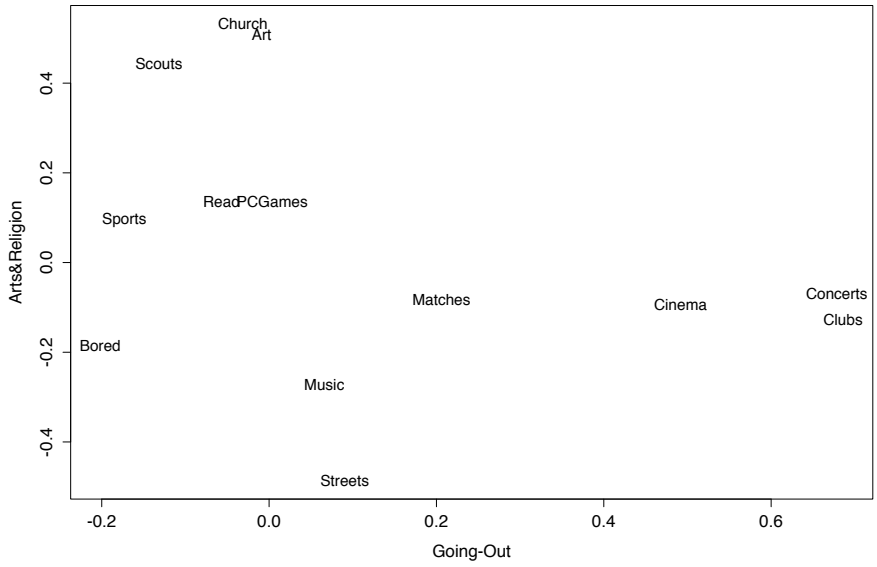

Figure 3.3: Leisure activities with respect to the first two factors.

"Music" as *"Msc"*, *"PC Games"* as *"PCG"*, *"Read"* as *"Red"*, *"Scouts"* as *"Sct"*, *"Sports"* as *"Spr"* and *"Streets"* as *"Str"*).

Now also the locations of the individuals can be examined in the metric space that has been constructed. The biplot in Figure 3.5 shows both location of the individuals with respect to the first two factors as well as the projections of the original axes of the leisure activities. This figure gives some intuition about the social space under consideration. The two factors *Going-out* and *Arts & Religion* appear to be quite excluding. Almost none of the pupils is strongly associated with both factors, i.e., those pupils that are going out a lot are very unlikely to spend time on arts and religion. However, the pupil with ID 84 in this picture forms an exception. He has the highest score for *Arts & Religion*, i.e., 2.36, and also a score clearly above the average for *Going-out*, i.e., 1.43. In Table 3.7 summary statistics of the factor loadings are displayed, i.e., mean, standard deviation (SD), minimum, maximum, maximum - minimum (range) and number of observations (N). Furthermore, as discussed in Section 2.2.2, one aim of geometric data analysis is to understand whether external factors help to explain the position of the individuals in the constructed social space. Here external factors are individual attributes that are not used when constructing the metric space, such as gender or age. To illustrate the idea, we use the gender of the pupils and their distributions across the different factors. Therefore we also include the summary statistics by gender into Table 3.7.

We clearly see that there are difference between the two genders regarding their average factor score. This is true for all five factors and the differences are significant as t-tests show. Females score on average significantly higher on the factor *Going-out* than male

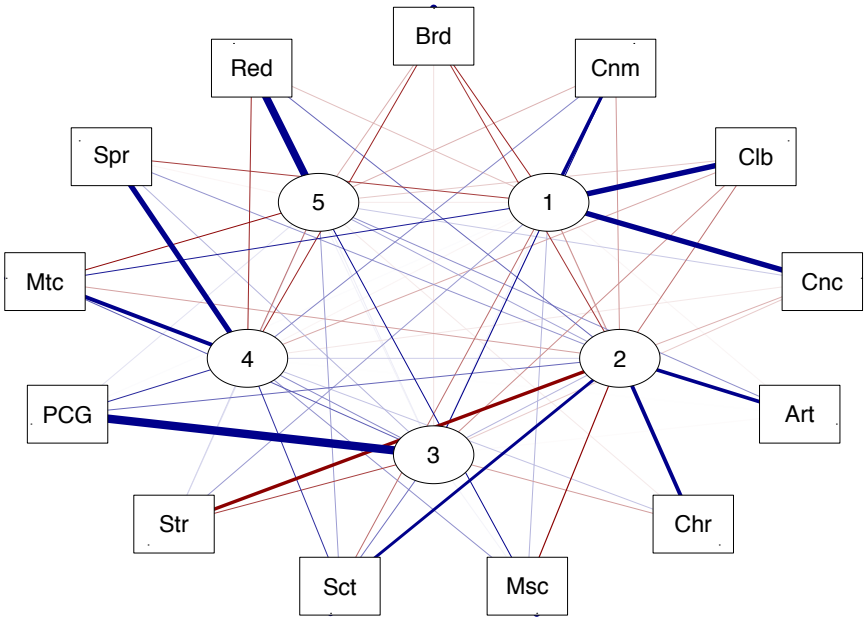

Figure 3.4: Associations between factors and variables. The numbers indicate Factor 1 to Factor 5. The labels are 3 characters abbreviations of the variable names.

pupils (0.16 vs. -0.14, $p < 0.05$). Females score on average significantly lower on the factor *Arts & Religion* than males (-0.16 vs. 0.15, $p < 0.05$). Female Pupils score on average significantly lower on the factor *Games & Movies* than males (-0.40 vs. 0.37, $p < 0.001$). Females score on average significantly lower on the factor *Sports* than males (-0.31 vs. 0.28, $p < 0.001$). On the other hand, females score significantly higher on the factor *Reading & Music* than males (0.28 vs. -0.25, $p < 0.001$). Thus, we clearly see that there are different behaviors related to females and males respectively. Particularly in this age, social norms might play a decisive role for a pupil when choosing a behavior. All three motives, compliance, identification and internalization, which we discussed in Section 2.1, will strongly impact the individuals to conform to gender roles and stereotype behavior. Thus, if we only know the gender of a pupil, we can already estimate the location of this pupil with respect to the identified factors.

Another aim of geometric data analysis is to understand whether the position of an individual in the social space explains individual behavior (see Section 2.2.2). As we are interested in the smoking behavior of the pupils, we will consider it with respect to the locations of the individuals. Therefore, we also break the summary statistics of the five factors down into non-smokers and smokers (see Table 3.7).

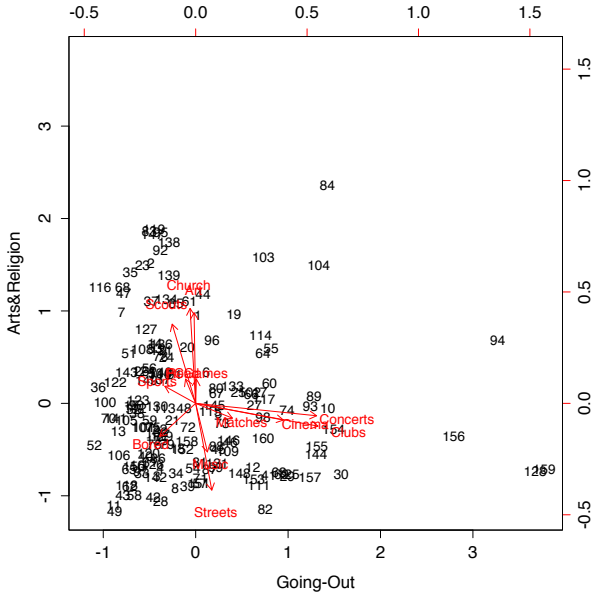

Figure 3.5: Position of the pupils with respect to the first two factors. Also the leisure activities are shown.

There are clear differences between smokers and non-smokers in their average factor scores; and for three out of the five factors these differences are significant as shown by t-tests, namely for *Arts & Religion*, *Games & Movies* and *Sports*. Smokers score on average significantly lower on the factor *Arts & Religion* than non-smokers (-0.35 vs. 0.10, $p < 0.001$). Smoking pupils also score on average significantly lower on the factor *Games & Movies* than non-smoking pupils (-0.34 vs. 0.10, $p < 0.001$). Furthermore, smokers score on average significantly lower on the factor *Sports* than non-smokers (-0.29 vs. 0.09, $p < 0.05$). For the other two factors, smokers score on average higher than non-smokers. However, here the differences are not significant.

In Figure 3.6 the location of the individuals with respect to all five factors is displayed with the help of a scatterplot matrix. Smoking pupils are colored red and non-smokers are colored blue. One can clearly see that subgroups of pupils are formed and that smokers are clustered together with respect to some of the factors.

Using clustering algorithms is one way to systematically identify *clouds of points* (see Section 2.2.2) so that we do not have to rely on visual inspection only. We apply k-means clustering [Mar14] to identify sub-groups of pupils. As a distance we use the Euclidean distance and based on the plot displayed in Figure 3.7, which can be regarded similar to a scree plot, we decide to extract four clusters.

In Table 3.8 the sizes of the resulting clusters are displayed. Furthermore, it is shown how high the pupils in each cluster score on average on each of the five factors. We see that

		Mean	SD	Minimum	Maximum	Range	N
Going-Out	All	0.00	0.83	-1.10	3.77	4.87	160
	Female	0.16	0.11	-1.10	3.77	4.87	76
	Male	-0.14	0.07	-1.05	3.27	4.32	84
	Smoking	0.24	0.14	-1.10	3.68	4.78	37
	Non Smoking	-0.07	0.07	-1.05	3.77	4.82	123
Arts & Religion	All	0.00	0.77	-1.17	2.36	3.53	160
	Female	-0.16	0.08	-1.14	1.88	3.03	76
	Male	0.15	0.09	-1.17	2.36	3.53	84
	Smoking	-0.35	0.09	-1.06	1.19	2.25	37
	Non Smoking	0.10	0.07	-1.17	2.36	3.53	123
Games & Movies	All	0.00	0.98	-1.61	1.62	3.23	160
	Female	-0.40	0.11	-1.47	1.58	3.05	76
	Male	0.37	0.10	-1.61	1.62	3.23	84
	Smoking	-0.34	0.16	-1.39	1.58	2.97	37
	Non Smoking	0.10	0.09	-1.61	1.62	3.23	123
Sports	All	0.00	0.77	-1.99	1.51	3.50	160
	Female	-0.31	0.08	-1.98	1.09	3.08	76
	Male	0.28	0.07	-1.63	1.51	3.15	84
	Smoking	-0.29	0.14	-1.98	1.24	322	37
	Non Smoking	0.09	0.07	-1.76	1.51	3.27	123
Reading & Music	All	0.00	0.91	-2.28	1.09	3.37	160
	Female	0.28	0.08	-1.73	1.09	2.82	76
	Male	-0.25	0.11	-2.28	1.09	3.37	84
	Smoking	0.12	0.14	-2.14	1.05	3.19	37
	Non Smoking	-0.04	0.08	-2.28	1.09	3.37	123

Table 3.7: Summary table of pupils' scores on the five factors (all pupils as well as broken down by gender and smoking behavior respectively).

in cluster 1 pupils are assembled that like to go out, to read an to listen to music.They do not spend their leisure time in the cinema or in front of the computer playing games. Also they do not watch sports or participate in sportive activities. Furthermore, they do not spend time on arts or an instrument and also not with religious activities. It is the biggest cluster comprising approximately one third of the pupils. Cluster 2 contains pupils that really like playing computer games and going to the movies. Furthermore, they spend their free time reading or listening to music. They do not pursue religious activities or dedicate their time to arts or and instrument. This cluster is the second largest comprising about 31% of the pupils. The pupils in cluster 3 really like sports, both watching sport matches and participating in sports. Furthermore, they like to spend their leisure time playing computer games and going to the cinema. They really dislike reading and listening to music. Furthermore, they do not play an instrument or dedicate their time to arts. Also they do not go to a church, mosque or temple. This cluster is the smallest cluster comprising 17.5% of the pupils. Finally, cluster 4 that usually dedicate their leisure time to arts or an instrument or to religious activities. Also sports

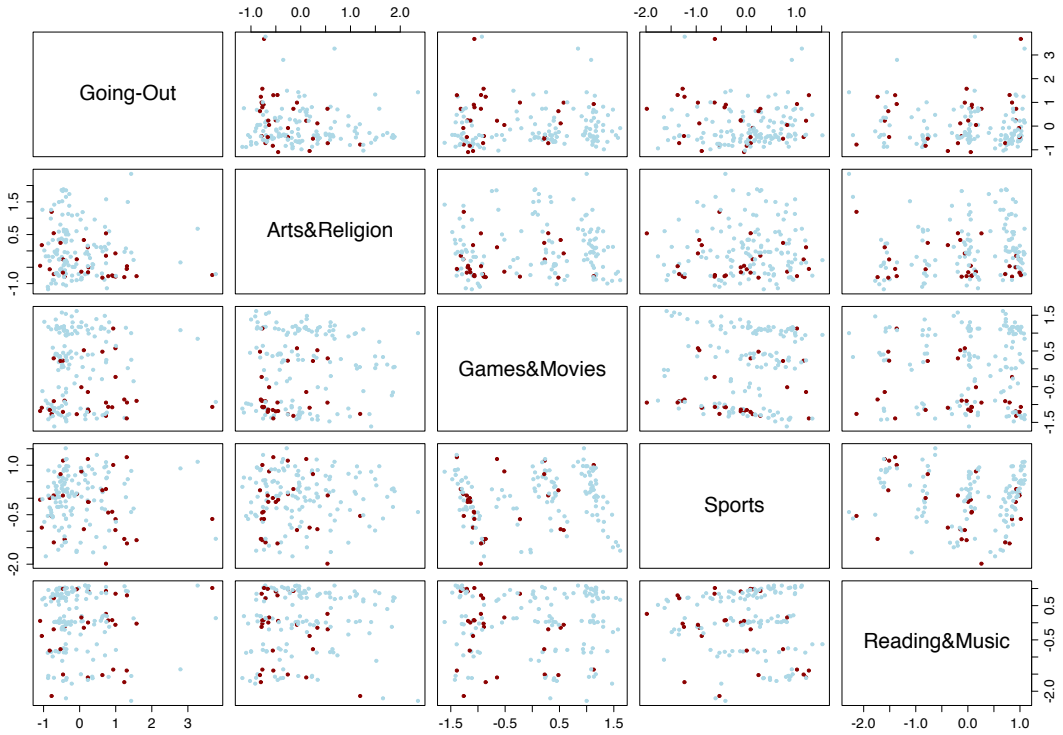

Figure 3.6: Scatterplot matrix of the five factors. Smokers are colored red, non-smokers are colored blue.

is interesting to them as well as reading and listening to music. They clearly do not like to go out. This cluster is the third largest comprising 18.1% of the pupils.

Now we also take a look at the identified clusters with respect to gender distribution and smoking behavior. In Table 3.9 the proportion of female pupils within each cluster is shown. Furthermore, the proportion of smokers in each of the clusters is given. Additionally, we list the proportion of smokers in each cluster at time t_3 to examine how smoking behavior evolves in the social system. Overall, the number of female and male pupils is quite balanced, there are 47.5% female pupils. However, if we look into the single clusters this proportion is varying a lot. Cluster 1 is almost exclusively comprising females, namely 86.8%. In all the other clusters, the proportion of female pupils is lower than 47.5% . In cluster 2 and cluster 4 there are 34 to 35% females. Cluster 3, on the other hand almost only comprises male pupils, here 10.7% are females. Thus, we again clearly see that the gender of a pupil already gives a lot of insights on behavioral patterns, i.e., external factors help to explain the position of the individuals in the social space.

Regarding the smoking behavior, there are 37 out of 160, i.e., 23.1%, smoking pupils at time t_2. Again, within the four clusters, this proportion varies considerably. The

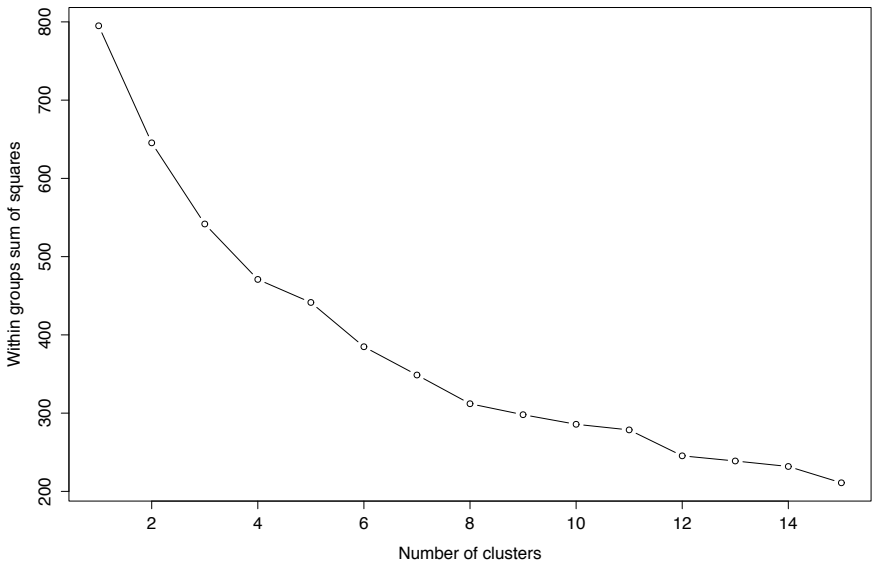

Figure 3.7: Plot to determine the number of clusters.

proportion of smokers in cluster 1 is much higher than the overall proportion, namely 43.4%. On the other hand, the proportion of smokers in cluster 3 and cluster 4 is very low, namely 7.1% and 6.9% respectively. The proportion of smokers in cluster 2 is a little bit below the average, namely 20.0%. Thus, different regions in the social space are related to different behavioral patterns. There is particularly one cluster in which smoking behavior is very common. This is valuable knowledge if someone would be interested in targeting the smoking pupils. Thus, the location in the social space indicates behavioral patterns.

At time t_3 overall a higher number of pupils smoke, namely 44 out of 160, i.e., 27.5%. Now, if we consider the single clusters, we see that also in each cluster the percentage of smoking pupils increases. In particular in cluster 1 approximately half of the pupils are smoking at time t_3. In cluster 3 and cluster 4 the proportion of smokers strongly increases due to the fact that there are only two smokers in each of those clusters at time t_2. However, at time t_3 there are four and three respectively. Thus, smoking is still not a representative behavior for these two clusters. In cluster 3 the proportion of smokers slightly increases as well. In general, the importance of smoking behavior in the single clusters does not considerably chance from time t_2 to time t_3. It slightly increases in all cluster in alignment with the overall development.

A closer inspection shows that from time t_2 to time t_3 13 pupils rather than seven start smoking. On the other hand six pupils quit smoking within this year. In particular, in cluster 1 behavioral changes occur. Here, eight pupils start and five quit smoking. In cluster 2 two start and one quits smoking. In cluster 3 two pupils start smoking and one

in cluster 4. The vast majority of pupils in cluster 1 is female. This is consistent with the results of the regression models indicating that there might be an association between being female and smoking (see Table 3.5).

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Size	53 (33.1%)	50 (31.3%)	28 (17.5%)	29 (18.1%)
Going-out	0.18 (1.21)	-0.04 (0.88)	0.08 (0.99)	-0.33 (0.70)
Arts & Religion	-0.48 (0.60)	-0.20 (0.63)	-0.43 (0.66)	1.64 (0.63)
Games & Movies	-1.00 (0.37)	0.96 (0.41)	0.22 (0.94)	-0.05 (0.83)
Sports	-0.58 (0.94)	-0.04 (0.96)	1.00 (0.56)	0.16 (0.69)
Reading & Music	0.24 (0.78)	0.50 (0.58)	-1.45 (0.44)	0.10 (1.06)

Table 3.8: Cluster sizes and standardized average scores (and standard deviations) of pupils belonging to this cluster with respect to the five factors.

	All	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Size	160 (100.0%)	53 (100.0%)	50 (100.0%)	28 (100.0%)	29 (100.0%)
Females	76 (47.5%)	46 (86.8%)	17 (34.0%)	3 (10.7%)	10 (34.5%)
Smokers t_2	37 (23.1%)	23 (43.4%)	10 (20.0%)	2 (7.1%)	2 (6.9%)
Smokers t_3	44 (27.5%)	26 (49.1%)	11 (22.0%)	4 (14.3%)	3 (10.3%)

Table 3.9: Number and proportion of female pupils and smokers at time t_2 and time t_3 in each cluster.

Furthermore, it is now possible use the insights regarding the different smoking behavior depending on the cluster a pupil belongs to when predicting this behavior. For example, we can include information about cluster membership within the logistic regression models. We use the previous model comprising the four prediction variables *female*, *relation*, *sibling smokes* and *drinking* with the outcome variable *smoking* at time t_2 as a basis for our discussion. It is listed again in Table 3.10. Now, when including the cluster membership as a predictor the model improves considerably. The Pseudo R^2 increases from 25.9% to 33.1%. In cluster 1 the highest percentage of smokers can be found, in cluster 2 the second highest, in cluster 3 the second lowest and in cluster 4 the lowest. This is captured by the coefficient of the predictor *cluster*. Being member of the next highest cluster leads to a 43% odds ratio to smoke at time t_2. Compared to Model 1, being female and the intercept term are not significant any more.

Now we try to predict the smoking behavior of the pupils at time t_3. Without the cluster membership, only drinking behavior is a significant predictor for that behavior. Furthermore, the intercept term is significant. Again, when including information about the cluster a pupil belongs to the model clearly improves; here the Pseudo R^2 increases from 24.6% to 30.3%. The predictor *cluster* is strongly significant. Being member of the next highest cluster leads to a 51% odds ratio to smoke at time t_3; the effect of smoking slightly decreases. Also the intercept is not significant any more (see Model 3 and Model 4 in Table 3.10).

We see that the leisure activities by themselves have not turned out to predict smoking behavior well (see Table 3.5). The clusters, on the other hand, capture the setting more accurately and have more predictive power. However, it is important to note again that the clusters are only build upon information on leisure activities.

	Dependent variable:			
	Smoking t_2		Smoking t_3	
	(1)	(2)	(3)	(4)
Female	0.719*	-0.081	0.603	-0.038
	(0.436)	(0.532)	(0.406)	(0.485)
Relation	0.847*	0.836*	0.647	0.629
	(0.449)	(0.472)	(0.435)	(0.451)
Sibling Smokes	0.892*	1.021*	0.545	0.627
	(0.502)	(0.533)	(0.496)	(0.513)
Drinking	1.418***	1.271***	1.615***	1.472***
	(0.423)	(0.438)	(0.402)	(0.412)
Cluster		-0.834***		-0.673***
		(0.305)		(0.261)
Constant	-2.568***	-0.497	-2.156***	-0.439
	(0.396)	(0.782)	(0.352)	(0.701)
Pseudo R^2	25.9%	33.1%	24.6%	30.3%
Log Likelihood	-71.495	-66.808	-79.189	-75.270
Akaike Inf. Crit.	152.991	145.616	168.378	162.539
N	160	160	160	160

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3.10: Logistic regression models taking into account cluster membership.

Summing up, with the help of the GDA approach we can assign behavioral patterns based on social context in a quite accurate way. The identified clusters are stable over time and help to analyze individual positions and behaviors. It is important to underline that neither information about the gender of the pupils nor about their smoking behavior is used to construct the metric space. Information about cluster membership can help to improve predictive models on the individual level, i.e., regression. Thus, it is a way to combine different levels.

Overall we are able to characterize the global picture very well. However, a more detailed analysis is necessary to capture the dynamics going on within the clusters. In Chapter 4, for example, we put focus on the pairwise interactions of the pupils and analyze how

these interactions impact smoking behavior. In Chapter 5 information on clusters and network structure are combined.

3.2 Collective Behavior and Recommmender Systems

Now our aim is it to apply the outlined approach to capture social context in the area of recommender systems. There it is crucial to understand user preferences and behaviors in order to deliver accurate recommendations. We will use the introduced approach to construct an abstract metric space based on observational data that contains both user profiles as well as products to recommend. Furthermore, we propose a novel, picture-based way to elicit user preferences and behavioral patterns in order to locate the user in the metric space. Here we will discuss our attempt in the context of travel recommender systems (based on [NSSW14a, NSSW14b]). However, our approach is generic and has meanwhile also been applied to other domains including events and jewelry.

3.2.1 Travel Preferences and Personality

To start the discussion, we refer to the characterization of travel preferences and behavior in literature. As discussed in Section 2.3.1, the 17 tourist roles identified in the work of Yiannakis and Gibson [YG92] have proved to be a good characterization of touristic behavioral patterns. However, up to now there was no attempt to reduce the number of tourist roles to a potentially lower number of factors that are also able to capture enough variance in touristic behavior. Although Berger et al. [BDD+07] mention in their work that some of the roles are very close, they do not pursue this question further. The work of Yiannakis and Gibson and similar approaches, moreover, mainly focus on what people are already doing but there is no attempt to predict what people would potentially enjoy.

Furthermore, although related work shows that travel behavioral patterns are not exclusive and are typically changing over a person's life course [GY02], none of these approaches makes any effort to describe a person's actual travel preferences as a combination of different travel aspects. However, this would make sense, in particular as related work suggests that people have a variation of travel preferences at the same time [GMHF04].

Our goal is, as opposed to this, to identify a lower number of tourist factors that are sufficient to capture enough variance in travel behavior. We introduce a model in which the travel profile of a user is composed of seven basic factors that can be used, moreover, to deliver travel related recommendations.

Moreover, to characterize user behavior and preferences more accurately, we also aim to take long-term behavior into account. Therefore, we include personality traits into our model because it has been shown that they allow to predict behavioral patterns over time and across situations [WRS02]. Therefore, we apply the so-called "Big Five" taxonomy. This taxonomy distinguishes five personality traits, i.e., *extraversion*, *agreeableness*, *conscientiousness*, *neuroticism*, and *openness to experience*. There is a huge body of psychological literature describing the theoretical background of the so-called Big Five

taxonomy and its emergence (e.g., [Gol93, JS99]). Despite some controversial aspects, the Big Five taxonomy is very overall well accepted and forms the basis of a large amount of work in various field including behavioral and consumer research. Importantly, personality traits exhibit longitudinal stability and are quite stable over a person's life course [Con85]. Furthermore, sophisticated questionnaires in many languages have been developed to assess a person's personality [Bor64]. There is also work that discusses how the Big Five taxonomy could be applied in the context of recommender systems. It has been shown that there is some association between music preferences and personality, which could be taken into account when delivering recommendations [HP10]. Furthermore, in [FST15] a relation between personality traits and picture properties such as color or saturation is shown.

Name	Abbreviation	Mean (SD)
Sun Lover	snl	3.69 (0.69)
Action Seeker	act	3.24 (0.73)
Anthropologist	ant	2.69 (1.06)
Archaeologist	arc	2.55 (0.71)
Organized Mass Tourist	omt	2.45 (0.73)
Thrill Seeker	trs	2.92 (0.8)
Explorer	exo	3.5 (0.66)
Jet Setter	jst	3.8 (0.64)
Seeker	skr	2.44 (0.81)
Independent Mass Tourist I	imt1	2.71 (1.07)
Independent Mass Tourist II	imt2	3.17 (1.03)
High Class Tourist	hct	2.18 (0.91)
Drifter	dtr	2.09 (0.91)
Escapist I	esc1	3.57 (0.9)
Escapist II	esc2	3.18 (1.04)
Sport Tourist	spt	3.12 (0.89)
Educational Tourist	edt	2.58 (0.97)
Openness	opn	3.91 (0.6)
Conscientiousness	cns	3.8 (0.58)
Extraversion	ext	3.25 (0.65)
Agreeableness	agr	3.18 (0.64)
Neuroticism	nrt	3.46 (0.52)

Table 3.11: Names and abbreviations of tourist roles and "Big Five" factors. Also mean and standard deviation of the normalized answers to the questionnaire are listed.

For our study, we develop a questionnaire considering both the 17 tourist roles by Gibson and Yiannakis and the Big Five traits. These 22 variables are listed in Table 3.11 togheher with the abbreviations that are used in the following discussion. The questionnaire comprises 50 questions, out of which 30 address the tourist roles (i.e., two questions per tourist role except for *Independent Mass Tourist 1*, *Independent Mass Tourist 2*, *Escapist 1* and *Escapist 2*; those are addressed by only one quesiton each) and 20 the personality traits (i.e., four questions per trait), in random order. Regarding the travel roles, we use the same questions as in the work of Gibson and Yiannakis [GY02]; the questions

addressing the personality traits are developed based on comprehensive documentations available on this subject. All questions are phrased as statements that are supposed to be rated on the basis of a five level scale, corresponding to the integers one (strongly disagree) to five (strongly agree). Additionally, we ask demographic information such as gender and age.

To collect the data, we made the questionnaire available offline and online. The offline questionnaires were filled out by randomly selected participants at crowded places in Vienna and in Austrian trans-regional trains. In this way, 553 forms were completed. The online questionnaire was promoted via Social Media. Here, additional information such as IP-addresses of the participants, starting time and overall time to complete the questionnaire was collected. This information was used to perform plausibility checks. In the end, we kept 444 completed online forms. Thus, in total we obtained 997 completed questionnaires.

Descriptive Statistics

Out of the 997 participants, who answered the questionnaire, 486 (48.7%) are female and 511 (51.3%) are male. In our sample, the majority is rather young. In Table 3.12 the age distribution of the participants in terms of six age groups is listed. As we see, the largest age group comprises the 20- to 29-year-olds (375 participants or 37.6%); and the second larges age group comprises the 30- to 39-year-olds (213 participants or 21.4%). The participants come from 59 different countries. However, most of them are from Austria (614 participants or 61.6%), followed by Slovakia (82 participants or 8.2%) and Germany (69 participants or 6.9%).

0 - 19	20 - 29	30 - 39	40 - 49	50 - 59	60 plus
82	375	213	139	113	75
8.2%	37.6%	21.4%	13.9%	11.3%	7.5%

Table 3.12: Age distribution of the participants.

As most of the tourist roles and all Big Five factors are addressed by more than one question, we aggregate and normalize the answers to those questions. Mean and standard deviation of the normalized answers are listed in Table 3.11. On average, the participants identify themselves most with *Openness*, *Jet Setter*, *Conscientiousness*, *Sun Lover*, and *Escapist I*. On the other hand, the participants identify themselves least with *Drifter*, *High Class Tourist*, *Seeker*, *Organized Mass Tourist* and *Archaeologist* (for the description of the tourist roles, see Table 2.1 on page 44).

In Figure 3.8 the correlations between the tourist roles and Big Five personality traits are displayed. The highest positive correlations occur between *Action Seeker* and *Anthropologist* (0.66, $p < 0.001$), *Independent Mass Tourist I* and *Seeker* (0.56, $p < 0.001$), *Sun Lover* and *Neuroticism* (0.46, $p < 0.001$), *Anthropologist* and *Archaeologist* (0.42, $p < 0.001$) as well as *Organized Mass Tourist* and *High Class Tourist* (0.40, $p < 0.001$). On the other hand, the highest negative correlations are between *Drifter* and

Conscientiousness (-0.25, $p < 0.001$), between *Sun Lover* and *Seeker* (-0.15, $p < 0.001$) and between *Sun Lover* and *Independent Mass Tourist I* (-0.12, $p < 0.001$).

In [DNW16] a detailed analysis of the distribution of the 22 variables across the different age groups and genders is presented. There it is shown that the popularity of the tourist roles strongly varies with age whereas the Big Five personality traits are quite equally distributed over the age groups. This confirms the long-term stability of personality traits. Furthermore, the distribution of tourist roles does not vary significantly between the genders.

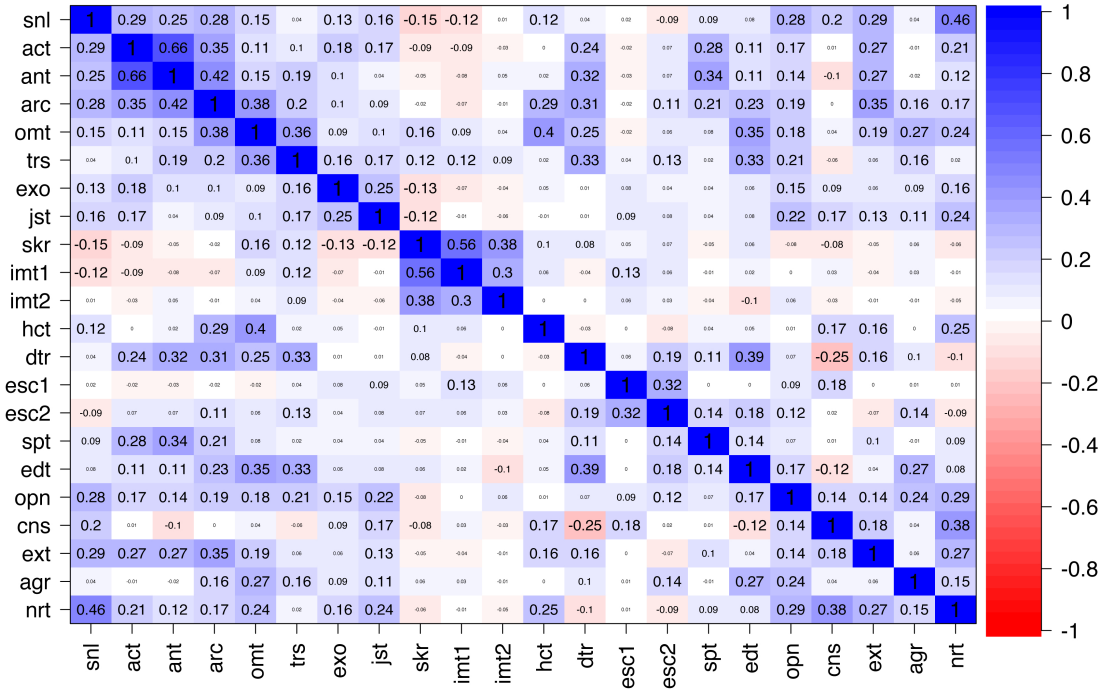

Figure 3.8: Visualization of the correlations between the 22 variables, i.e., 17 tourist roles and five personality traits.

Factor Analysis

Based on this data we perform a factor analysis with basically the same steps as before (see Section 3.1.3). According to the KMO criterion, the data is suitable for factor analysis; the overall MSA is 0.7 and each variable has a MSA of at least 0.59. In the scree plot in Figure 3.9), we can see that six eigenvalues are larger than one and one eigenvalue is approximately equal to one. After testing results with different numbers of factors, we decide to keep a seven factor solution with varimax rotation. These factors meet our requirements: 1) the seven factors account for a sufficiently large amount of variance, namely 58%; and 2) it is possible to identify adequate interpretations for the

seven factors. As we want to apply the results in the context of recommender systems, we consider the latter even more important; it enables to communicate the results to the users in a sound and understandable way.

In Table 3.13 the resulting factor loadings that are bigger than 0.20 or smaller than -0.20 are displayed. We see that the obtained model represents the data well and is meaningful: The first factor is highly associated with the tourist role *Sun Lover* (loading 0.50) and the personality trait *Neuroticism* (loading 0.79). Also strong associations with *Openness* (loading 0.51) and *Conscientiousness* (loading 0.53) can be found. We summarize these characteristics as *Sun & Chill-Out*. The second factor combines the personality trait *Agreeableness* (loading 0.46) with the tourist roles *Organized Mass Tourist* (loading 0.63), *Drifter* (loading 0.51) and *Educational Tourist* (loading 0.67). We call this aspect *Knowledge & Travel*. The third factor is strongly associated with the tourist roles *Seeker* (loading 0.78), *Independent Mass Tourist I* (loading 0.68) and *Independent Mass Tourist II* (loading 0.49). To summarize those properties, we describe this aspect as *Independence & History*. Since the fourth factor combines *Archaeologist* (loading 0.81) and *High-Class Tourist* (loading 0.59) with the personality trait *Extraversion* (loading 0.41), we call it *Culture & Indulgence*. The fifth factor represents the tourist roles *Sun Lover* (loading 0.28), *Anthropologist* (loading 0.73), *Drifter* (loading 0.37) and *Sport Tourist* (loading 0.36) as well as the personality trait *Extraversion* (loading 0.28). We characterize it as *Social & Sport*. A combination of *Action Seeker* (loading 0.40), *Explorer* (loading 0.51) and *Jet Setter* (loading 0.51) yields the sixth factor, which we summarize as *Action & Fun*. The tourist roles *Escapist I* (loading 0.43) and *Escapist II* (loading 0.72) form our last factor, which we summarize as *Nature & Recreation*.

Thus, we are able to reduce the number of variables considerably. Starting from 22 variables, i.e., 17 tourist roles plus five personality traits, we determined seven basic factors that form the basis of an abstract metric space that we will use to model users and products. We can visualize the tourist roles and the Big Five factors in the constructed space. In Figure 3.10 all 22 variables that have been considered in the factor analysis are located with respect to the dimensions *Sun & Chill-Out* and *Knowledge & Travel*. Variables that are located along a dimension strongly characterize this dimension. Such pictures help to interpret the setting. In Figure 3.11 the associations between the 22 variables and the seven factors are visualized. Positive loadings are shown in blue, negative loadings are shown in red. The higher the loading the thicker and darker the connection line.

In order to check whether both approaches for data collection, i.e., offline and online, lead to similar results, we conduct the factor analysis three times; first we use the offline questionnaires only, second the online questionnaires only and third we use both combined. The identified factors appear to be robust, there are no essential differences for the different approaches.

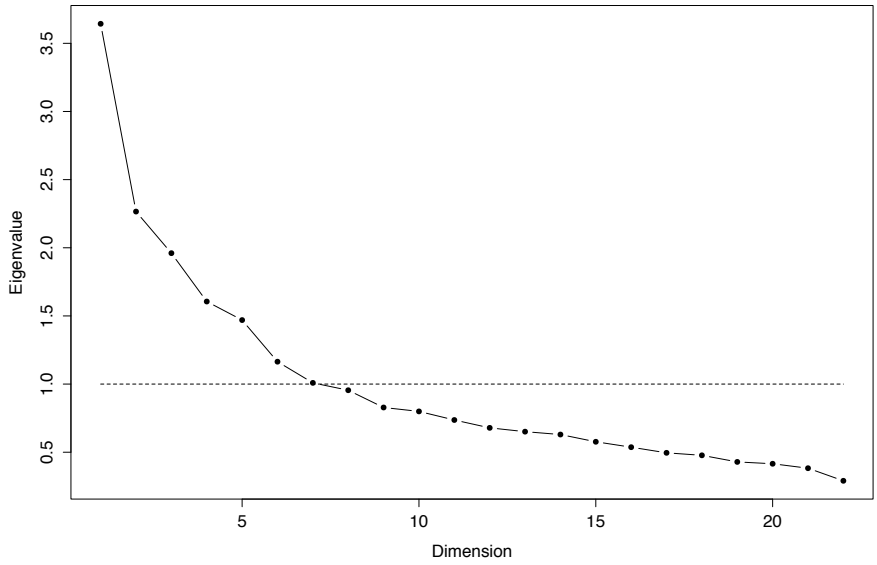

Figure 3.9: Tourist roles and Big Five factors: scree plot.

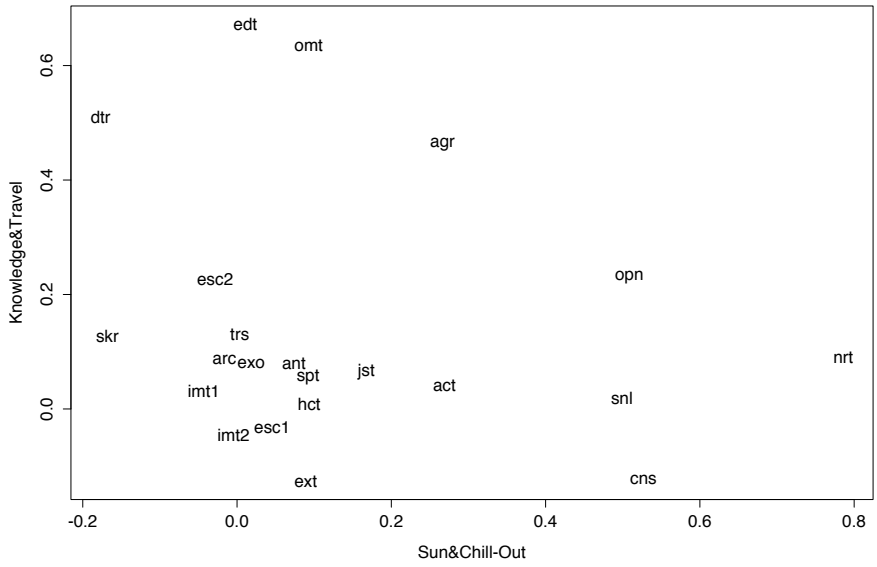

Figure 3.10: Tourist roles and personality traits with respect to the first two factors.

	Sun & Chill-Out	Knowledge & Travel	Independence & History	Culture & Indulgence	Social & Sport	Action & Fun	Nature & Recreation
snl	0.50			0.25	0.28		
act	0.27					0.40	
ant					0.73		
arc				0.81			
omt		0.63					
trs							
exo			-0.21			0.51	
jst						0.51	
skr			0.78				
imt1			0.68				
imt2			0.49				
hct				0.59			
dtr		0.51			0.37		
esc1							0.43
esc2		0.23					0.72
spt					0.36		
edt		0.67					
opn	0.51	0.23					
cns	0.53						
ext				0.41	0.28		
agr	0.27	0.46		-0.22			
nrt	0.79						

Table 3.13: Factor analysis: loadings of the seven factor solution (only the loadings greater than 0.20 or smaller than -0.20 are displayed).

3.2.2 Using Pictures for Preferences Elicitation

In line with GDA (see Section 2.2.2), we aim to explore the preferences of the individuals based on their location in the constructed space. However, we want to apply this model in the context of recommender systems, where it is crucial to accurately locate the users with respect to the basic factors. In order to meet this challenge we propose a novel way for preference elicitation, i.e., using pictures.

In order to identify suitable travel related pictures and to establish an association between them and the seven factors, we pre-select and pre-process 102 pictures and assign each of those pictures to one of the factors. This was done in an interdisciplinary workshop with participants from technical disciplines, tourism as well as artistic background. The participants had the task to consider the iconic value of an image and thus to abstract from the actual picture. Thus, an opera represents the concepts music and culture rather than showing the Vienna State Opera. Furthermore, the participants of the workshop had to reach a consensus on the assignments of the pictures to the factors.

After that, a second study was conducted with 105 new participants. The study has the following set-up:

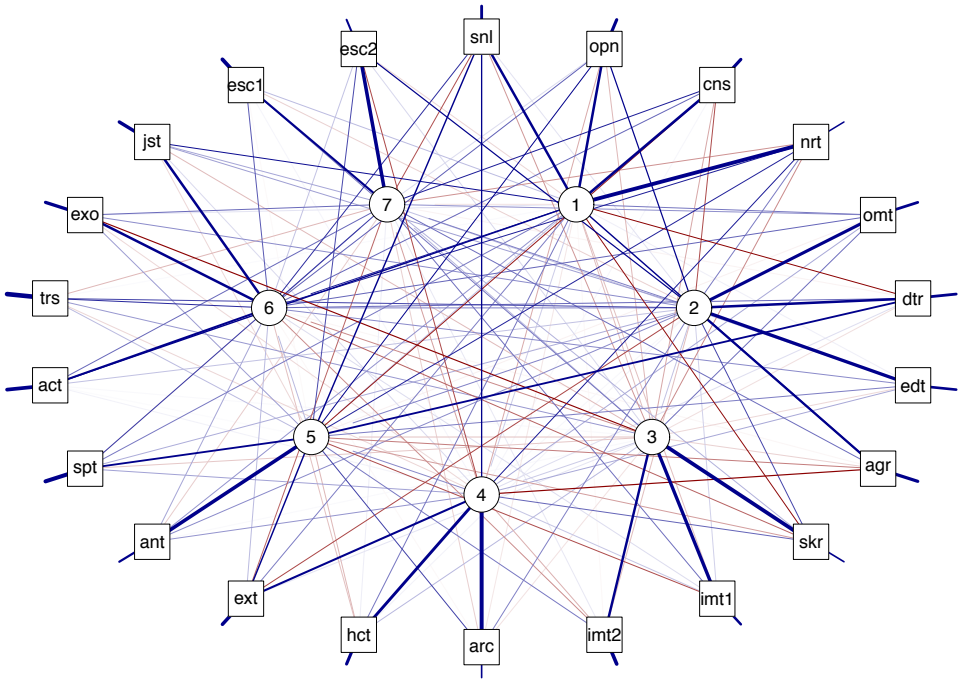

Figure 3.11: Associations between the 22 variables and the seven factors.

- As illustrated in Figure 3.12 (left-hand side) we arrange the pictures on a table in random order and in rectangular formation.
- Then each participant is asked to select three to ten pictures that she finds most appealing when thinking about her next hypothetical vacation.
- The participants are given the opportunity to make an intermediate step, i.e., to make a pre-selection of pictures before taking a final decision. Furthermore, they can also put pictures back.
- Based on this pre-selection, the final picture set has to be chosen. Furthermore, the participants are asked to rank the selected pictures.
- Finally, the participants are also asked to fill out the questionnaire used in the previous study (see Section 3.2.1). This helps to validate the assignment of the pictures to the factors and to test the robustness of the identified seven-factor solution.

This study leads to following insights: First, people typically select between three and seven pictures. Second, the intermediate step is not required, people tend to add pictures immediately to the final selection rather than to the pre-selection. Third, the order of the pictures in the final selection is usually reconsidered and changed. Fourth, not all

pictures have the same popularity. Therefore, we remove on the one hand the pictures that have never been selected and on the other hand the pictures that are almost always selected. In that way, we are able to reduce the initial set of pictures to 63. For these 63 pictures, moreover, the following applies: first, all of them are used to identify one ore more factors; and second, each factor is related to a number of pictures.

Figure 3.12: Picture selection – offline (left-hand side) and online (right-hand side) [NSSW14a]

.

Now we also want to place items that can be recommended to the users in the metric space. Therefore, we select 10,835 points of interest (POI) all over the world. This selection is based on an online investigation for popular tourism places and activities in various countries. The POIs are retrieved from open source tourism data bases or we purchase commercial licenses. We categorize the POIs either as *Sight* (e.g., the Eiffel Tower), *Activity* (e.g., boat ride), *Restaurant*, *Entertainment* (e.g., an opera or musical), *Shopping*, *Nightlife* or *Tour*. To place these touristic items in the metric space and to compute the model for recommendation, we collaborate with 15 tourism experts who are frequent travelers and know the respective POIs. They receive the task to assign to each POI between three and seven representative pictures out of the picture set. Furthermore, they have to rank the assigned pictures according to their degree of association with the POI. We ask them, moreover, to assign a number from the interval $[0, 1]$ for each POI and each factor. This numbers are supposed to express the extent to which in the opinion of the expert a traveler associated with a certain factor would enjoy the respective POI. Thus, the experts have to provide seven numbers for each POI, i.e., one per factor.

Now we were able to quantify the associations between the travel related factors and the pictures. Separately for each of the seven factors a multiple regression model of the form

$$F_j = b_{j0} + \sum_{i=1}^{63} b_{ji} x_i, \tag{3.1}$$

$j = 1, \dots 7$, is developed. However, before being able to calculate the coefficients b_{ji}, we have to use the assignments of the tourism experts first: For the respective factor, we set the left-hand side of the equation F_j to the assigned number from the interval $[0, 1]$. To identify adequate values for x_i, $i = 1, \dots, 63$, we test different approaches:

1. Here, applying a so-called "dummy coding", we set x_i to 1 if the i-th picture has been selected for the respective factor and to 0 otherwise. In this case, we do not consider the information about the ranking.

2. In this approach, we are taking the ranking provided by the tourism experts into account. We set x_i to $(-k + 8)/7$ if the i-th pictures has been selected and ranked on the k-th place. Thus, the values corresponded to 1 for the first ranked picture, 6/7 for the second one, 5/7 for the third one, and so on.

3. Here, we set
$$x_i = 7\frac{-k + n + 1}{\sum_{i=1}^{n} i} \tag{3.2}$$
if the i-th picture has been selected and ranked on the k-th place and 0 otherwise. In this case, x_i does not only depend on rank k but also on the number of selected pictures n. If, for example five pictures are selected, x_i is equal to 35/15 for the first ranked picture, 28/15 for the second one, and so on. Furthermore, the numbers x_i are always adding up to seven.

After assigning values to the x_i, we are now able to estimate the coefficients b_{ji} with the help of the ordinary least squares method. In order to avoid over-specification, the intercepts b_{j0} are set to 0. To compare the three approaches, we develop several test cases. Then, based on those test cases, we assess the results of the three approaches with respect to their accuracy and plausibility. Overall, the third method performs best. The first approach is too simple and the second one turns out to be highly dependent on the size of the chosen subset.

The third approach forms the basis of our recommender algorithm. It also allows us to determine the travel profile of a user. After she chooses between three and seven pictures, the x_i are computed as described previously. Thus, let x_i^u, $i = 1, \dots, 63$ be the computed values x_i for user u and b_{ji}, $i = 0, \dots, 63$, $j = 1, ..., 7$ the coefficients of the seven equations of the model. Then we obtain the seven factors F_j^u, $j = 1, \dots, 7$ for the user by

$$F_j^u = \sum_{i=1}^{63} b_{ji} x_i^u, \tag{3.3}$$

The model is designed in a way that the resulting F_j^u are in the range $[0, 1]$ quantifying how much a user is represented by each factor. For example, if $F_1^u = 0.6$, $F_2^u = 0.3$, $F_3^u = 0.4$, and so on, user u is a mix of travel factor 1 (60%), travel factor 2 (30%), travel

factor 3 (40%), and so on. As the equations are independent, the percentages typically exceed 100%. Also the ranked order of the selected pictures has an impact on a user's travel profile.

Since now the travel profile of a user is known, we can determine in a next step her points of interest. In order to obtain the scores F_j^p of the factors F_j, $j = 1, \ldots, 7$ for a POI p, we consider the values from the interval $[0, 1]$ that have been assigned by the tourism experts (as described previously) and aggregate them in an appropriate way. In line with GDA, we understand the factors as a basis of a seven dimensional vector space. Thus, each POI represents also a point in this space. As also the users are located in the same metric space, we can compute the distances between the travel profile of a user and different POIs. For this purpose we apply the Euclidean metric $d : \mathbb{R}^7 \times \mathbb{R}^7 \to \mathbb{R}$,

$$d(F^u, F^p) = \sqrt{\sum_{j=1}^{7} (F_j^u - F_j^p)^2}, \tag{3.4}$$

where vector F^u is representing the travel profile of user u and vector F^p is representing the POI p in the seven-dimensional space. The algorithm will now recommend those POIs to the user that are closest to her travel profile with respect to the Euclidean metric.

	F_1	F_2	F_3	F_4	F_5	F_6	F_7	Distance
User profile	0.14	0.70	0.79	0.88	0.30	0.20	0.12	
Stonehenge	0.09	0.74	0.75	0.89	0.21	0.03	0.17	0.213
Daibutsu	0.06	0.76	0.82	0.82	0.38	0.03	0.07	0.229
Wat Maheyong	0.06	0.65	0.80	0.84	0.35	0.00	0.11	0.231

Table 3.14: User profile and recommended POIs.

As example we present a user profile with its respective F_j scores (between 0 and 1) and the first three recommendations: The profile $F_1 = 0.14$ (*Sun & Chill-Out*), $F_2 = 0.70$ (*Knowledge & Travel*), $F_3 = 0.79$ (*Independence & History*), $F_4 = 0.88$ (*Culture & Indulgence*), $F_5 = 0.30$ (*Social & Sport*), $F_6 = 0.20$ (*Action & Fun*), and $F_7 = 0.12$ (*Nature & Recreation*) leads to the recommendations: *Stonehenge* (a prehistoric monument in England with the scores 0.09, 0.74, 0.75, 0.89, 0.21, 0.03, 0.17), *Daibutsu* (a temple in Japan with the scores 0.06, 0.76, 0.82, 0.82, 0.38, 0.03, 0.07), and *Wat Maheyong* (a Buddhist temple in Thailand with the scores 0.06, 0.65, 0.80, 0.84, 0.35, 0.0, 0.11). Applying the Euclidean metric, we can compute the distances between the user profile and the POIs. In Table 3.14 we see that Stonehenge is closer to the user profile than the other two POIs.

This approach has been implemented and forms the core of the picture-based search and recommendation engine PixMeAway [Pix14]. In Figure 3.12, right-hand side the interface for online picture selection is shown. It constitutes the first step of the customer's interaction with the system. The individual travel profile of the user is presented as a

feedback with the help of comic characters; it is shown to what extend the preferences of the user align with each of those travel factors (see Figure 3.13). In a next step (not shown) users can indicate their preferences, such as time or point of departure. Then, the recommendations are displayed.

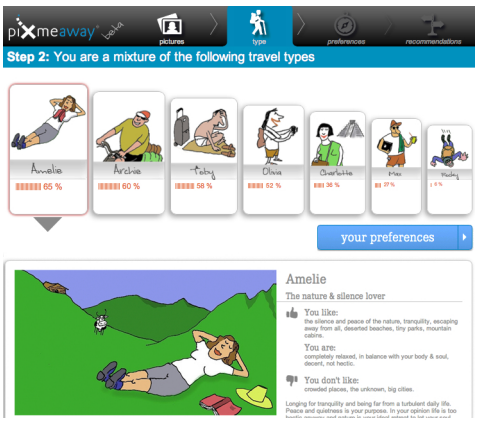

Figure 3.13: User interface – travel profile feedback [NSSW14a].

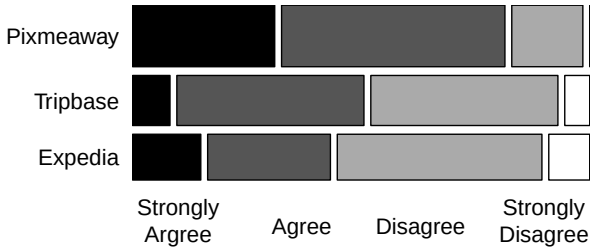

Figure 3.14: Mosaic plot: "The page is exciting" (based on [Kri12]).

In a first user study, aspects of the preference eliciting interface with respect to user liking were investigated [Kri12]. In total, 110 participants were assigned to one of the following Websites in order to explore them: PixMeAway [Pix14], Tripbase, a travel Website providing also recommendations [Tri14], and Expedia, a traveling booking site [Exp14]. After that, their experiences with respect to several emotional categories, such as inspiration, enjoyment, enthusiasm and interest were inquired using an online questionnaire. The main goal of this work was to study the impact of emotional and inspirational aspects of online travel portals on the overall satisfaction of users and their intention to revisit a Website rather than doing a "classical" evaluation measuring precision and recall (since at this stage users have no clear idea what they want).

The findings of this first analysis are promising. For the category excitement, the Kruskal-Wallis test detected significant differences between the distributions of the answers regarding the Websites ($p < 0.001$). Here, 84% of the participants agreed or strongly

agreed that the experience with PixMeAway was exciting (see Figure 3.14). There are similar results for the categories inspiration, interest, enjoyment and pride. Furthermore, the study showed a significant correlation between inspiration and the intention to revisit a Website ($p < 0.01$) as well as between inspiration and satisfaction ($p < 0.001$).

3.2.3 User Positions and Travel Factors

Individual Attributes

Having constructed the metric space of preferences and travel behavioral patterns, we can now analyze whether characteristics such as age and gender can explain the positions of the users with respect to the seven travel factors.

First we focus on the different age groups. In Table 3.15 the summary statistics of user scores for the seven factors are listed for all users as well as broken down by age groups. When comparing the average factor scores across different age groups, we find clear differences for most of the factors.

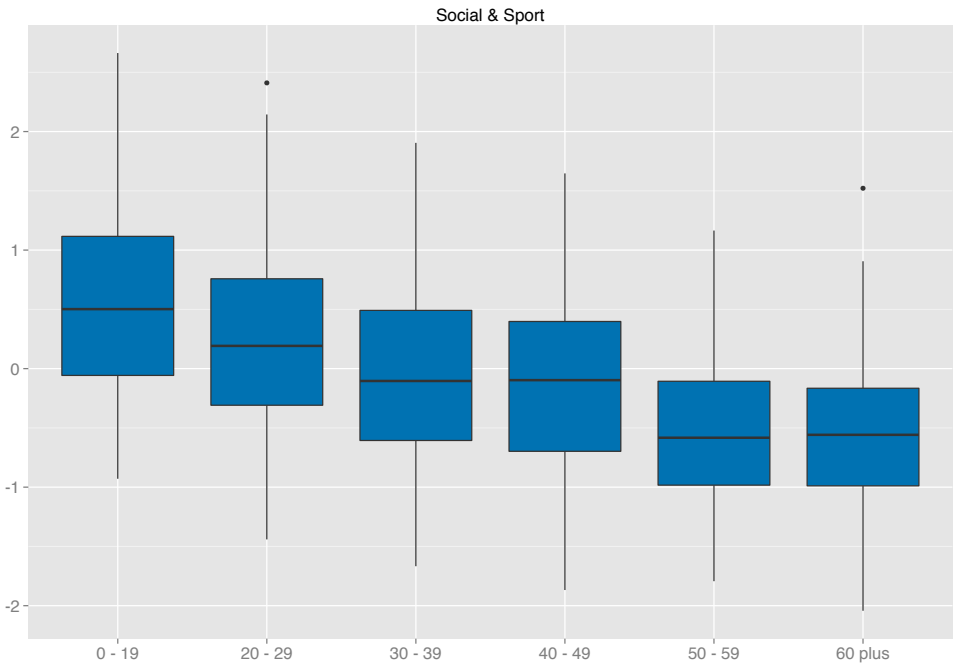

Figure 3.15: Score distributions with respect to *Social & Sport* of the different age groups.

		Mean	SD	Minimum	Maximum	Range	N
Sun &	All	0.00	0.03	-3.89	2.54	6.43	997
Chill-Out	0 - 19	-0.15	0.09	-2.26	1.46	3.73	82
	20 - 29	-0.07	0.04	-3.89	2.54	6.43	375
	30 - 39	0.16	0.06	-2.73	2.01	4.74	213
	40 - 49	0.02	0.08	-3.35	1.91	5.26	139
	50 - 59	-0.04	0.08	-2.03	2.30	4.33	113
	60 plus	0.06	0.10	-2.23	2.06	4.29	75
Knowledge	All	0.00	0.03	-2.51	2.71	5.23	997
& Travel	0 - 19	-0.22	0.09	-2.51	2.04	4.55	82
	20 - 29	0.13	0.05	-2.02	2.71	4.73	375
	30 - 39	-0.04	0.06	-1.83	2.47	4.31	213
	40 - 49	-0.20	0.06	-1.97	1.72	3.68	139
	50 - 59	-0.02	0.07	-1.94	2.03	3.98	113
	60 plus	0.12	0.08	-1.62	1.66	3.28	75
Independence	All	0.00	0.03	-2.14	2.70	4.84	997
& History	0 - 19	0.05	0.10	-1.88	2.10	3.98	82
	20 - 29	0.06	0.04	-2.14	2.70	4.84	375
	30 - 39	-0.03	0.06	-1.92	2.50	4.41	213
	40 - 49	-0.14	0.07	-1.72	2.27	3.99	139
	50 - 59	-0.01	0.08	-1.93	2.13	4.06	113
	60 plus	0.02	0.10	-1.73	2.58	4.32	75
Culture &	All	0.00	0.03	-1.67	3.02	4.69	997
Indulgence	0 - 19	0.30	0.10	-1.10	2.73	3.83	82
	20 - 29	-0.02	0.05	-1.41	3.02	4.43	375
	30 - 39	-0.02	0.06	-1.21	2.66	3.88	213
	40 - 49	-0.14	0.06	-1.33	1.69	3.02	139
	50 - 59	-0.02	0.07	-1.67	2.28	3.95	113
	60 plus	0.10	0.10	-1.43	2.64	4.07	75
Social &	All	0.00	0.03	-2.04	2.66	4.71	997
Sport	0 - 19	0.49	0.09	-0.93	2.66	3.59	82
	20 - 29	0.24	0.04	-1.44	2.41	3.85	375
	30 - 39	-0.05	0.05	-1.67	1.90	3.57	213
	40 - 49	-0.13	0.07	-1.87	1.65	3.51	139
	50 - 59	-0.53	0.06	-1.79	1.17	2.96	113
	60 plus	-0.55	0.08	-2.04	1.52	3.56	75
Action	All	0.00	0.02	-3.35	2.01	5.36	997
& Fun	0 - 19	0.03	0.08	-1.98	1.61	3.59	82
	20 - 29	0.10	0.04	-3.35	1.76	5.11	375
	30 - 39	0.04	0.04	-1.88	1.65	3.52	213
	40 - 49	-0.22	0.06	-2.29	1.19	3.48	139
	50 - 59	-0.11	0.07	-2.37	2.01	4.38	113
	60 plus	-0.05	0.09	-2.38	1.51	3.89	75
Nature &	All	0.00	0.02	-2.32	1.92	4.24	997
Recreation	0 - 19	-0.36	0.11	-2.29	1.66	3.95	82
	20 - 29	-0.08	0.04	-2.32	1.68	4.00	375
	30 - 39	0.11	0.05	-2.07	1.87	3.94	213
	40 - 49	0.15	0.06	-1.67	1.76	3.43	139
	50 - 59	0.20	0.08	-1.80	1.92	3.72	113
	60 plus	-0.09	0.08	-1.87	1.68	3.55	75

Table 3.15: Summary table of user scores on the seven factors (all users as well as broken down by age group).

Pairwise t-tests with p-value adjustment show that there are statistically significant differences in the means for all factors but *Independence & History*. In particular, with respect to *Sun & Chill-Out* the age group 20 to 29 with an average of -0.07 scores significantly lower than the age group 30 to 39 with an average of 0.16. With respect to *Knowledge & Travel*, the age group 20 - 29 with an average of 0.13 scores significantly higher than the age group 0 - 19 with an average of -0.22 as well as the age group 40 - 49 with an average of -0.20. With respect to the factor *Culture & Indulgence*, the age group 0 - 19 scores significantly higher than the age group 20 - 29 and 40 - 49 with an average of -0.02 and -0.14 respectively. However, *Social & Sport* shows clear and significant differentiations between most of the age group (there are no significant differences between the age groups 30 - 39 and 40 - 49 and between the age groups 50 - 59 and 60 plus). The age group comprising the youngest users has the highest average score, namely 0.49. The scores are then decreasing; the older the users the lower their average factor score (see also Figure 3.15).

With respect to *Action & Fun*, the age group 40 - 49 with an average of -0.22 scores significantly lower than the age groups 20 - 29 with an average score of 0.10 and 30 - 39 with an average score of 0.04. Also *Nature & Recreation* exhibits differences between the age groups. The age group 0 - 19 with an average of -0.36 scores significantly lower than all other age groups but 60 plus. Furthermore, there are also significant differences between the age group 20 - 29 with an average of -0.08 to all the other age groups but 60 plus. We can also see that in the youngest age group, i.e., beyond 20 years, the average scores for *Sun & Chill-Out*, *Knowledge & Travel* and *Nature & Recreation* are below the respective average in the entire sample. On the other hand, members of this age group score on average high with respect to *Culture & Indulgence* and *Social & Sport*. People in the largest age group, i.e., between 20 and 29 years, score comparably low with respect to *Sun & Chill-Out* and *Nature & Recreation*. On the other hand, people in their twenties score on average high with respect to *Knowledge & Travel* and *Social & Sport*. People between 30 and 39 years typically enjoy *Sun & Chill-Out* as well as *Nature & Recreation*. People in the age 40 to 49 years like on average *Nature & Recreation* and score low with respect to all the other factors except for *Sun & Chill-Out*. People between 50 and 59 years also score on average high with respect to *Nature & Recreation*. On the other hand, they score negative with respect to *Action & Fun* and clearly dislike *Social & Sport*. Also people in the age group 60 plus typically dislike *Social & Sport*. They also score on average negative with respect to *Nature & Recreation* and *Action & Fun*.

Now we consider gender. In Table 3.16 summary statistics of user scores per gender are listed for female and male users. Here, the differences are significant for *Sun & Chill-Out*, where females score on average higher than male users, *Independence & History*, where also females score on average higher and *Social & Sport*, where males score on average much higher than females.

In Section 3.2.1 we introduce the questionnaire on the 17 tourist roles and Big Five personality traits that serves as a basis for the factor analysis. As we discuss there, some tourist roles are more popular among the participants than others and some personality

		Mean	SD	Minimum	Maximum	Range	N
Sun &	Female	0.11	0.04	-3.70	2.54	6.24	486
Chill-Out	Male	-0.10	0.04	-3.89	2.30	6.20	511
Knowledge	Female	-0.05	0.04	-2.51	2.45	4.97	486
& Travel	Male	0.05	0.04	-2.02	2.71	4.73	511
Independence	Female	0.08	0.04	-1.93	2.70	4.63	486
& History	Male	-0.07	0.04	-2.14	2.59	4.73	511
Culture &	Female	-0.01	0.04	-1.41	3.02	4.43	486
Indulgence	Male	0.01	0.04	-1.67	2.66	4.34	511
Social &	Female	-0.16	0.04	-2.04	2.66	4.71	486
Sport	Male	0.15	0.04	-2.00	2.41	4.41	511
Action	Female	0.04	0.03	-2.38	1.76	4.14	486
& Fun	Male	-0.03	0.03	-3.35	2.01	5.36	511
Nature &	Female	-0.04	0.04	-2.32	1.92	4.24	486
Recreation	Male	0.04	0.03	-2.28	1.76	4.05	511

Table 3.16: Summary table of user scores on the seven factors by gender.

traits occur more frequently. In [DNW16] it is shown that the popularity of the tourist roles strongly varies between the age groups but not between the genders. Now we have been discussing the same characteristics, i.e., age groups and gender, with respect to the seven latent dimensions. This turns out to be quite useful as the differentiations become more distinctive. We are able to detect clear differences between all age groups and genders with respect to factors scores. Less dimensions are easier to handle. Furthermore, by using latent structure, noise in the data can be reduced.

Knowing about the different positions of the users in the metric space dependent on their age or gender is clearly an advantage in the context of recommender systems. Information about these characteristics is often available and thus can be efficiently used for personalization and targeting marketing as we can relate the characteristics to travel preferences and behavioral patterns.

Groups of Users

In a next step we use the identified seven dimensions to determine groups of people who share similar travel preferences and behavioral patterns (*"clouds of points"*). Here, we apply the same approach as in Section 3.1.3. In Figure 3.16 a plot is shown, which helps to determine the number of clusters. Based on this, we try different solutions. In the end, we decide to work with six cluster also because of their interpretability. In Table 3.17 the cluster means with respect to the factors are displayed.

We see that cluster 1 comprises people who like *Action & Fun* as well as *Nature & Recreation*. Furthermore, *Sun & Chill-Out* is important to them. In their holidays, this people do not look for aspects related to *Independence & History* or *Knowledge & Travel*. Also *Culture & Indulgence* is not important to them. For people belonging to cluster

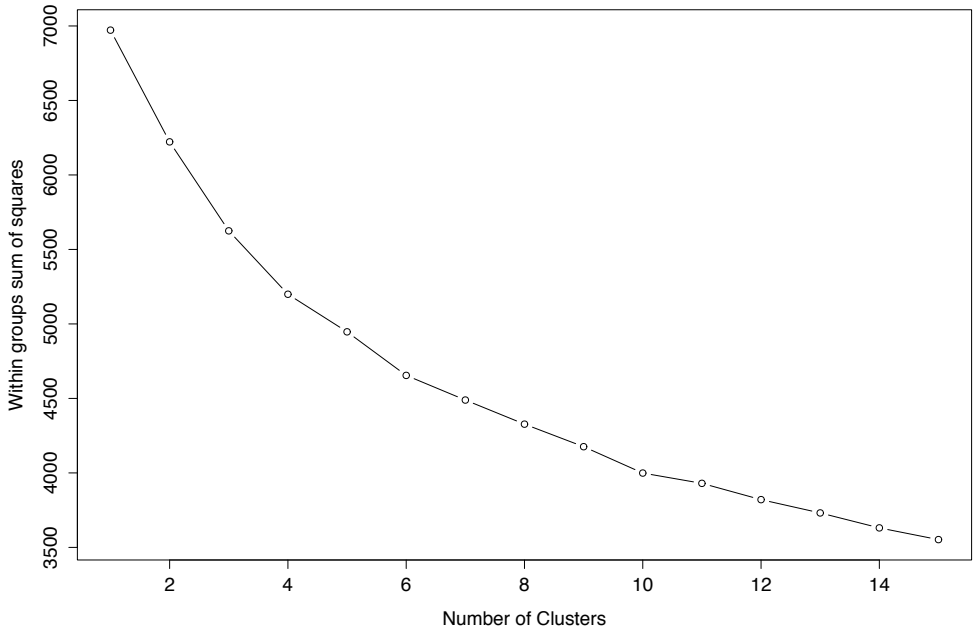

Figure 3.16: Shared travel preferences: plot to determine the number of different groups.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Sun & Chill-Out	0.40	0.57	-0.06	-1.10	0.26	-0.19
Knowledge & Travel	-0.52	0.57	0.98	-0.18	-0.46	-0.36
Independence & History	-0.62	1.08	-0.29	0.60	-0.18	-0.39
Culture & Indulgence	-0.37	0.60	-0.44	-0.50	-0.32	1.46
Social & Sport	-0.10	-0.05	0.83	-0.31	-0.66	0.34
Action & Fun	0.68	0.18	0.34	-0.83	-0.31	-0.25
Nature & Recreation	0.68	0.43	-0.08	0.49	-0.97	-0.66
N	197	150	179	159	177	135

Table 3.17: Description of the clusters with respect to the seven factors (the mean scores of the users by factor in each of the clusters are displayed).

2, on the other hand, *Independence & History* is very important. Moreover, they like *Knowledge & Travel* and *Sun & Chill-Out* as well as *Nature & Recreation*.

The most important aspects for people in cluster 3 are *Knowledge & Travel* as well as *Social & Sport*. Furthermore, they appreciate *Action & Fun*. Those people are not interested in *Culture & Indulgence* or *Independence & History*. People in cluster 4 have interest in *Independence & History* and *Nature & Recreation*. They do not want to spend their holidays with *Sun & Chill-Out* or *Action & Fun*. Also *Culture & Indulgence* is not of interest. In cluster 5 people are pooled together who basically disapprove to all aspects

but *Sun & Chill-Out*. Cluster 6, finally, comprises people who particularly like *Culture & Indulgence*. Also *Social & Sport* is important to them. On the other hand, they do not like *Nature & Recreation*. Statistical tests (i.e., pairwise t-test with p-value adjustment) clearly show that there are significant differences between the clusters with respect to the different factors. Thus, people belonging to a certain cluster or group exhibit a distinct travel behavior compared to people belonging to the other clusters.

Now we examine how age and gender are distributed among the identified clusters. In Table 3.18 it is clearly visible that there are differences between the clusters regarding age. In cluster 1 the three age groups 30 - 39, 40 - 49 and 50 -59 year olds are over-represented compared to the overall population. In cluster 2, the 30 - 39 year olds are clearly over-represented. They constitute 28% whereas in the overall population they only constitute 21.4% of the people. Also the 50 - 59 year olds and the age group 60 plus are slightly over-represented. Cluster 3 mainly comprises people in their twenties; 58.1% of the people in the cluster belong to this age group whereas in the overall population people of this age only constitute 37.6% . In cluster 4 the age groups 40 - 49 and 50 - 59 are clearly over-represented; the former constitutes 21.4% of this cluster (compared to 13.9% in the overall population) and the latter 17.6% (compared to 11.3% in the overall population). In cluster 5 people who are 40 years and more are over-represented. In particular, the age group 60 plus comprises 15.3% of this cluster whereas in the overall population they only constitute 7.5%. Finally, in cluster 6 the 0 - 19 year olds are strongly over-represented. In the overall population they only account for 8.2% but in cluster 6 for 19.3%. Also the 40 - 49 year olds are over-represented in that cluster but only slightly.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	All
All	197	150	179	159	177	135	997
	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)
0 - 19	11	11	13	10	11	26	82
	(5.6%)	(7.3%)	(7.3%)	(6.3%)	(6.2%)	(19.3%)	(8.2%)
20 - 29	66	54	104	49	55	47	375
	(33.5%)	(36%)	(58.1%)	(30.8%)	(31.1%)	(34.8%)	(37.6%)
30 - 39	51	42	39	27	29	25	213
	(25.9%)	(28%)	(21.8%)	(17%)	(16.4%)	(18.5%)	(21.4%)
40 - 49	32	9	13	34	29	22	139
	(16.2%)	(6%)	(7.3%)	(21.4%)	(16.4%)	(16.3%)	(13.9%)
50 - 59	26	21	6	28	26	6	113
	(13.2%)	(14%)	(3.4%)	(17.6%)	(14.7%)	(4.4%)	(11.3%)
60 plus	11	13	4	11	27	9	75
	(5.6%)	(8.7%)	(2.2%)	(6.9%)	(15.3%)	(6.7%)	(7.5%)

Table 3.18: Distribution of the different age groups in the different clusters.

In Table 3.19 the gender distribution across the clusters is displayed. In the overall sample, there are 48.7% females and 51.3% males. Women are clearly over-represented in cluster 2 (58%) and in cluster 5 (55.9%) and slightly over-represented in cluster 1

(49.7%). Men, on the other hand are clearly over-represented in cluster 3 (60.9%) and cluster 4 (58.5%). In cluster 6 the female/male distribution strongly resembles the gender distribution in the overall sample.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	All
Female	98	87	70	66	99	66	486
	(49.7%)	(58%)	(39.1%)	(41.5%)	(55.9%)	(48.9%)	(48.7%)
Male	99	63	109	93	78	69	511
	(50.3%)	(42%)	(60.9%)	(58.5%)	(44.1%)	(51.1%)	(51.3%)

Table 3.19: Distribution of female and male users in the different clusters.

Summing up, we have used travel related behavior and preferences of individuals to construct an abstract metric space. The dimensions of this space are determined by the latent behavior of the people. This facilitates the description of the social system and allows for analyzing the behavior of the people in this space. We clearly see that people belonging to different groups behave differently. If the users that are modeled represent customers these groups can be interpreted as customer segments. In our case we want to recommend tourism products with the help of pictures (as discussed in Section 3.2.2). Since the pictures enable us to position a user in the metric space, they also can be used to predict the group a user belongs to. We are applying this approach in a related project where we develop models to predict eight customer segments based on pictures. There, preliminary results show that we are able to predict the correct customer segment of a user with a rather high accuracy (around 60%). For the travel domain a similar study is planned.

What we have discussed here can be subsumed as comparison mechanism. Most of the users will not know each other and might never have interacted. However, their behaviors are similar and predictable. Here, clearly social norms play a crucial role.

3.3 Discussion

In this chapter we discuss comparison based social influence mechanism on the group level. Applying the GDA approach we construct metric spaces based on collective behavior and preferences. First we demonstrate the concepts with the help of a data set on teenage smoking behavior from the literature. Conducting a factor analysis on the leisure time behavior of the teenagers leads to the five dimensions *Going-out*, *Arts & Religion*, *Games & Movies*, *Sports* and *Reading & Music*. Based on these factors, which constitute a space of lifestyle, we can demonstrate two main assumptions of GDA, i.e., external variable such as gender (i.e., variables that are not used when constructing the space) are able to explain the position of the individuals in the space and the position of the individuals in the space, in turn, indicates their behavior (i.e., smoking). We conduct a cluster analysis leading to a four clusters solution. The identified clusters are stable over time and characterize the setting in an accurate way. In particular, we include information

about cluster membership in logistic regression models and show that this is a good predictor for smoking behavior.

In a next step we apply the same approach to picture-based recommender systems. We developed a computational model of travelers based on the seven basic factors *Sun & Chill-Out*, *Knowledge & Travel*, *Independence & History*, *Culture & Indulgence*, *Social & Sport*, *Action & Fun* and *Nature & Recreation*. These factors capture latent travel behaviors and preferences and are obtained by a factor analysis of the 17 tourist roles and the Big Five personality traits. This approach can be regarded as a content and a knowledge based recommender approach.

Furthermore, the travel experience is typically emotional. Therefore, we enable the users to express their preferences with the help of pictures. The idea is to provide decision support in a stage of the travel decision-making process where users are not able to express their needs explicitly. This results in a new approach to elicit preferences where a user does not have to verbalize her interests nor translate them into product attributes. Furthermore, there is no need for a long interaction with the system.

In this work, we examine how well the seven factors, which consider both individual short and term behavior, differentiate between distinct groups of users. Statistical analyses show that that there are significant differences between the identified groups. It becomes apparent that age and gender play a crucial role for individual travel behavior. Thus, travel behavior is not only constituted by personality traits but also by social norms and normative behavior. Additional insights could be gained by also considering detailed interactions among the users, e.g., related to communication based influence. This will be done in future work.

CHAPTER 4

Social Influence in Networks

In Chapter 3 we have introduced a geometric approach how to incorporate social context in models of human behavior. Goal of this approach is to identify dimensions of latent behavior that form the basis of a metric space so that the locations of the individuals in that space allow to draw conclusions about individual preferences and emerging behavior (see also Section 2.2.2). With the help of this approach we are able to characterize the global picture quite well. However, so far all individuals have been treated as independent, we have not taken interactions between them into consideration. This will particularly be addressed in the following discussion where we examine communication-based social influence processes on the network level taking into account fine-grained information on dyadic relations among the individuals. Furthermore, we will again compare the insights that can be gained on this level with the individual level. Thus, with respect to Figure 4.1, we are here focusing on Part A and Part C of the figure.

We will use the example of teenage smoking behavior again to illustrate and compare different ways to address communication based social influence in a network setting. In this context we will propose a new way to capture such mechanisms. After that, we conduct an empirical study using the introduced models to analyze churn behavior in a multiplayer online game. Finally, we discuss the interdependencies of user sentiments in an online travel forum.

4.1 Communication-based Social Influence

Social influence occurs when individuals adapt their behavior (or attitudes, beliefs, etc.) according to the behavior of their friends. Now our aim is to model and understand explicitly the impact of the social network on individual behavior.

We focus on cross-sectional models rather than temporal ones. Furthermore, we now assume that a link in the network is either present or absent, i.e., the relations are treated

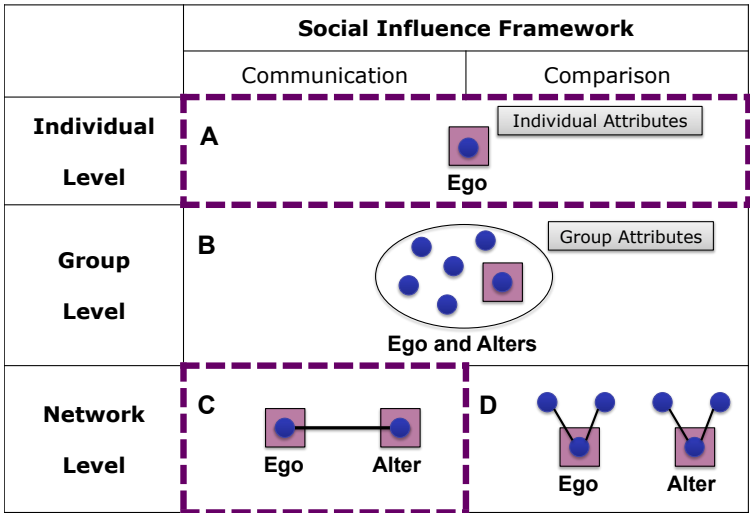

Figure 4.1: Social influence processes and levels of information.

as binary. Furthermore, we first focus also on binary behavior, i.e., an individual either displays a behavior or not.

In Section 2.2.4 we discuss the social influence models *Linear Network Autocorrelation Models* (LNAMs) and *Autologistic Actor Attribute Models* (ALAAMs). Furthermore, we introduce *Conditional Random Fields Models* (CRF Models) as a novel way to capture social influence in networks. Here we aim to compare these models to logistic regression models (Logit Models), introduced in Section 2.2.1. We summarize the different approaches briefly before starting with the analysis (see also Table 4.1):

The equation

$$\text{odds}(\Pr(\text{Y} = \text{y}|\text{X} = \text{x})) = \text{e}^{\sum \beta x} \tag{4.1}$$

describes the logistic regression approach. Here, the impact of individual attributes on the probability of the dependent behavior is independent. In this context, this is a clear limitation since this implies that logistic regresission is not capable of taking structure into account. Thus, also social influence cannot be captured. However, we aim to study data that is clearly not independent but interconnected, there is a high probability that the logistic regression models are mis-specified. As it is well-known in statistics this can lead to wrong estimations based on unreliable standard errors [FCS10]. Nevertheless logistic regression is often used with network data since it can handle big data sets and a high number of variables. Parameters are estimated by the maximum likelihood approach. In our study we use R software [R F16] to estimate logistic regression models.

The equation

$$y = \rho W y + \sum \beta x \tag{4.2}$$

captures LNAMs. Here, the behavior of a person is influenced by the weighted linear combination of the neighbors' behaviors. LNAM can be seen as an extension of OLS

regression for networks. Again, maximum likelihood is used for parameter estimation. However, we are using the model here for a binary outcome variable, which might also cause some issues with the standard error. Another limitation of this model is matrix W as the weighting scheme is critical for the outcome. One has to be aware of this when assigning weights to the edges in the network. However, here we only consider whether an edge exists or not. The computational costs are medium high. Here we use the R package *sna* [Car10], which contains a function *lnam*, to study this approach.

ALAAMs can be described by the equation

$$\Pr(\mathrm{Y} = \mathrm{y}|\mathrm{G} = \mathrm{g}, \mathrm{X} = \mathrm{x}) = \frac{1}{k} \cdot \mathrm{e}^{\sum \theta \cdot z(y,g,x).} \tag{4.3}$$

Here, the distribution of behavior is studied across network ties; the joint probability of network and behavior is modeled. A contagion parameter captures the number of neighbors of an individual that have the behavior under consideration; and its impact on the probability that the individual also displays this behavior quantifies the contagion effect. The network is undirected and weights cannot be assigned to the edges. The parameters are estimated by MCMC methods. As these models are very complex, the computational costs are considerably high and there are scalability issues. To develop ALAAMs we use the iPnet application [WRP06].

Finally we apply CRF models. These models are captured by the equation

$$\Pr(\mathrm{Y} = \mathrm{y}|\mathrm{X} = \mathrm{x}) = \frac{1}{Z} \cdot \mathrm{e}^{\sum \lambda \cdot f(e,y) + \sum \mu \cdot g(v,y,x)} \tag{4.4}$$

Here the network is regarded as undirected and fixed; the distribution of behavior is modeled in this fixed network. The impact of edge related parameters on the probability of the dependent, binary behavior is independent of the impact of node related attributes. These models are widely used in computer science, e.g., natural language processing, computer vision, bioinformatics. We propose to apply them to model social influence in networks; there are various advantages. CRF Models are able to capture interdependencies between the nodes. As opposed to ALAAM, also parameters can be assigned to different combinations of behavior, i.e., not only the impact of the behavior under consideration but also the impact of the other behavior can be captured by a separate parameter. Loopy Belief Propagation is used for parameter estimation. The computational costs are quite low, these models are scalable for more variables, more nodes and different types of networks. As a software, we use Matlab/UGM [Sch16] and adapt it to our problem. In particular, we add code to compute the standard errors and the p-values for the coefficients μ of the nodes. This is done straightforwardly, i.e., based on the covariance matrix, the standard errors and in turn the p-values are computed. For the edge parameters λ, we use a simulation approach to determine their significance. Based on the network under consideration we shuffle the behavior and assign it randomly to the nodes; i.e., the numbers of nodes that display the behavior and those who don't are fixed but they are linked differently. Then we count the pair of nodes that both exhibit the studied behavior. This is repeated k times, where k is a sufficiently large number. Based

on the distribution that is obtained by this procedure we can compute the expected value of pairs of nodes that both display the behavior as well as their standard deviation. This we can compare to the observed value in order to compute its p-value. To determine the standard errors for the parameters λ will be part of future work.

	Relations	Peer Influence	Estimation Method	Computational Costs	Software
Logit	No	No	MLE	Low	R/glm
LNAM	Weighted	Weighted sum of peers' behavior	MLE	Medium	R/lnam
ALAAM	Binary (undirected)	Contagion (Probability of peers with positive behavior)	MCMC	High	iPnet
CRF	Binary (undirected)	Probability of peers with any combination of behavior	LBP	Low	Matlab/ UGM

Table 4.1: Social Influence in networks: summary of methods.

4.1.1 Glasgow Data Set and Friendship

Also for this analysis we use the Glasgow data set from the *Teenage Friends and Lifestyle Study*, a study that aims to identify processes that lead to changes in the smoking behavior of teenagers in their early to mid adolescence in Glasgow (see Section 3.1).

In this longitudinal study, data on demographic characteristics and other individual attributes (e.g., being in a romantic relationship) as well as on behavior (e.g., smoking, drinking) and friendship was recorded at three time points (i.e., in 1995, 1996, and 1997); the sample consisted of 160 pupils in a secondary school.

To construct the network between the pupils, we consider their friendship connection. At each time point they were asked to indicate who of the other pupils is their friend, their best friend or no friend. We take this information from time point t_1 and create an undirected edge between two pupils if they both consider each other as a friend or at least one considers the other as a best friend. We apply this criterion to make sure that the tie between the pupils is a strong tie. The resulting network contains 399 friendship ties between the pupils. We consider the individual characteristics of the pupils at time point t_2 as in Section 3.1. We list in Table again descriptive statistics of the variables female, romantic relation, sibling smokes and drinking broken down by smoking behavior as we need this information for the following discussion. We see that overall 23.1% of the pupils are smoking. However, this percentage is considerably exceeded by female pupils, with a fraction of smokers of 32.9%; by pupils that are in a romantic relationship, with a fraction of smokers of 41.5%; by pupils who have at least one smoking sibling,

with a fraction of smokers of 44.4%; as well as by pupils who are drinking alcohol, with a fraction of smokers of 45.7%. Further descriptive statistics can be found in Section 3.1.

	Smoking	Non-Smoking	All
All	37	123	160
	(23.1%)	(76.9%)	(100%)
Female	25	51	76
	(32.9%)	(67.1%)	(100%)
Male	12	72	84
	(14.3%)	(85.7%)	(100%)
Relation	17	24	41
	(41.5%)	(58.5%)	(100%)
No Relation	20	99	119
	(16.8%)	(83.2%)	(100%)
Sibling smokes	12	15	27
	(44.4%)	(55.6%)	(100%)
No sibling smokes	25	108	133
	(18.8%)	(81.2%)	(100%)
Drinking	21	25	46
	(45.7%)	(54.3%)	(100%)
Non-Drinking	16	98	114
	(14.0%)	(86.0%)	(100%)

Table 4.2: Distribution of attributes with respect to smoking behavior.

Thus, we consider the friendship network of the pupils at time t_1 and their attributes and behaviors at time t_2. The reason is that we clearly want to separate these aspects to prevent our models to capture strong social selection effects rather than social influence (see also Section 2.2.3). In Figure 4.3 the friendship network between the pupils is displayed. Here, also some further attributes are visualized, i.e., the different colors capture smoking behavior (i.e., smokers are purple, non-smokers are orange) and the node size captures the gender of the pupil (i.e., females are represented by bigger nodes than males). Out of the 399 edges, 260 (65.2%) are between non-smokers, 99 (24.8%) represent a friendship tie between a non-smoker and a smoker and 40 (10.0%) connect two smokers.

4.1.2 Analysis and Results

To compare the four models, i.e., Logit, LNAM, ALAAM and CRF, we use smoking behavior as outcome variable and being female, being in a romantic relation, having a smoking sibling as well as drinking alcohol as predictor variables (see also Logit Model 2 in Table 3.5 on page 62). Furthermore, we include the friendship network of the pupils in the last three models, i.e., those models that are able to take structure into account. As the sample size is rather small and the effects are not so strong, we choose as a significance threshold a p-value of 0.1. This allows us to observe the differences

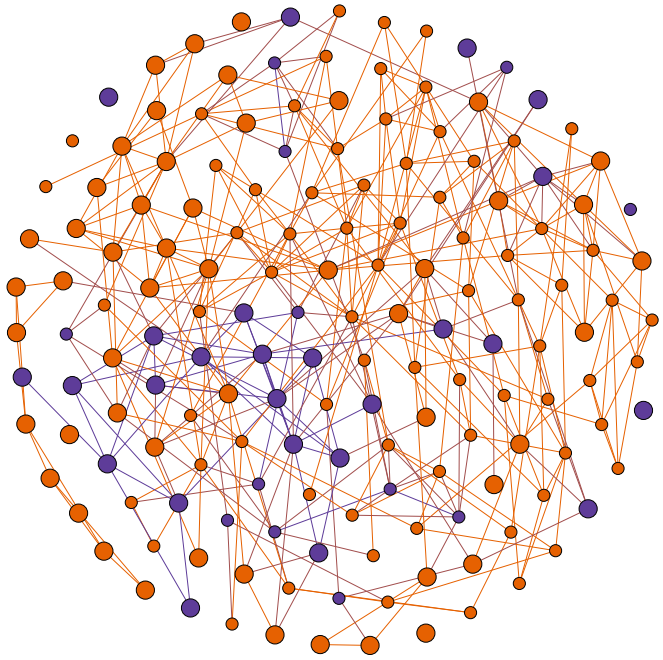

Figure 4.2: Friendship and smoking (smokers are displayed in purple, non-smokers in orange; nodes that represent females are bigger).

between the models better, which is the main focus of the discussion. The results for the four models are displayed in Table 4.3.

The predictor female is only significant in the Logit model, and also here it is only marginally significant. However, as soon as relations are taken into consideration, this effect disappears completely. Interestingly, male pupils are stronger connected than female pupils; males have on average 5.19 friends and females 4.78. However, if we consider smoking friends, the pictures changes; male pupils have on average 0.73 smoking friends whereas females exceed this number clearly by having on average 1.55 smoking friends. Thus, it could be the case that social influence through network ties causes the effect rather than gender. Being in a romantic relationship has a significant impact on smoking behavior in the ALAAM. In all the other models this predictor is marginally significant. Having a smoking sibling is clearly significant in the CRF Model and marginally significant in the Logit Model as well as the LNAM. In the ALAAM it is not significant at all. Thus, here the results are quite diverse. However, it is known in the literature that smoking siblings have a high impact on teenage smoking behavior (see [VS13]). Thus, the result of the CRF Model is reasonable. Drinking is clearly significant in all four models and it has the highest coefficient in all of them. The network effect is strongly significant in all the three models that are capable of considering structure. The contagion parameter assesses the smoking behavior of network neighbors; it captures the probability that a pupil is smoking given that friends are smoking. In the LNAM this effect is

	Dependent variable:			
	Smoking			
	Logit	LNAM	ALAAM	CRF
Constant	-2.568***		-1.983***	-3.033***
	(0.396)		(0.537)	(0.398)
Female	0.719*	0.075	0.054	0.331
	(0.436)	(0.055)	(0.374)	(0.443)
Relation	0.847*	0.156*	0.933**	0.894*
	(0.449)	(0.068)	(0.445)	(0.459)
Sibling smokes	0.892*	0.162*	0.847	1.007**
	(0.502)	(0.077)	(0.530)	(0.510)
Drinking	1.418***	0.221***	1.288***	1.320***
	(0.423)	(0.064)	(0.440)	(0.434)
Contagion		0.054***	0.755***	0.448***
		(0.016)	(0.182)	
Activity			-0.252**	
			(0.101)	
N	160	160	160	160

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4.3: Smoking behavior: results of model comparison.

considered to be linear, i.e., the more of the network neighbors are smoking the higher the impact. However, for ALAAM and CRF Models this effect is more sophisticated and it cannot be interpreted that straightforwardly. The contagion parameter and also the other coefficients are smaller in the LNAM compared to the other models but there is no intercept in the LNAM. The Logit Model, ALAAM and CRF have significant negative intercepts or constant terms, i.e., the basic probability of smoking is quite low. In the ALAAM additionally the activity of a pupil is included, which is significantly negative. It implies that pupils who are more active, i.e., who are well connected to other pupils regardless of their smoking behavior, are less likely to smoke.

4.1.3 Predictive Power of the Models

Now we also want to compare the predictive power of the models. Therefore, we use the fitted values of each model to predict both the pupils' non-smoking behavior as well as their smoking behavior.

First we assess the Goodness-of-Fit of the different models by comparing the predicted values with the observed values, i.e., the smoking behavior at time t_2. We consider the *accuracy*, the *precision*, the *recall* as well as the F_1 measure, which combines precision and recall:

$$\text{Accuracy} = \frac{\text{Total Number of Correct Predictions}}{\text{Number of Pupils}} \tag{4.5}$$

$$\text{Precision} = \frac{\text{Number of Correct Predictions}}{\text{Number of Predictions}} \tag{4.6}$$

$$\text{Recall} = \frac{\text{Number of Correct Predictions}}{\text{Number of (Non}-\text{)Smokers}} \tag{4.7}$$

$$\text{F}_1 \text{ Measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.8}$$

At time t_2 123 pupils are non-smoking and 37 are smoking. The Logit Model predicts non-smoking correctly in 118 cases and smoking in 11 cases, i.e., the accuracy equals 80.6%. The LNAM predicts non-smoking correctly in 118 cases and smoking in 13 cases, i.e., the accuracy equals 81.9%.

With ALAAM the prediction does not work straightforwardly as we cannot retrieve a probability for each node to display a certain behavior. Here, we have to do a simulation based on the estimated parameter. Our procedure is the following: we simulate 1000 networks and then assign to a node smoking behavior if the node displayed this behavior in a sufficiently high number of simulations. However, since it turns out that hardly ever the same nodes display smoking behavior in different simulations, we choose a low cut-off threshold of 2.7%, i.e., if a node displays the behavior in at least 27 out of the 1000 simulations, we assign this behavior to the node. With this procedure, ALAAM predicts non-smoking correctly in 104 cases and smoking in 3 cases, i.e., the accuracy equals 66.9%. Finally, the CRF Model predicts non-smoking correctly in 119 cases and smoking in 13 cases, i.e., the accuracy equals 82.5%.

In Table 4.4 the precision, recall and F_1 measures are listed. We see that also here CRF Model perform best. However, except for ALAAM the predictive power of the models is quite close, in particular LNAM has the same recall for predicting smoking behavior as the CRF Model.

Next we predict the non-smoking and smoking behavior respectively for the following year, assuming that there could be a delayed influence. In this year, 19 pupils change their smoking behavior, six quit smoking and thirteen start smoking. However, at time t_3 116 pupils are non-smoking and 44 are smoking. The Logit Model predicts non-smoking correctly in 111 cases and smoking in 11 cases, i.e., the accuracy equals 76.3%. The LNAM predicts non-smoking correctly in 109 cases and smoking in 11 cases, i.e., the accuracy equals 75.0%. Again, using a cut-off threshold of 2.7%, the ALAAM predicts

	Non-Smoking			*Smoking*		
	Precision	Recall	F_1	Precision	Recall	F_1
Logit	0.819	0.959	0.883	0.688	0.297	0.415
LNAM	0.831	0.959	0.890	0.722	**0.351**	0.472
ALAAM	0.754	0.846	0.797	0.136	0.081	0.102
CRF	**0.832**	**0.967**	**0.894**	**0.765**	**0.351**	**0.481**

Table 4.4: Goodness of Fit of the four models.

non-smoking correctly in 98 cases and smoking in 4 cases, i.e., the accuracy equals 63.8%. Finally, the CRF Model predicts non-smoking correctly in 111 cases and smoking in 12 cases, i.e., the accuracy equals 76.9%.

In Table 4.5 the precision, recall and F_1 measures for the prediction of the behavior at t_3 are listed. Again, the CRF Model perform best and ALAAM worst. The predictive power of the models are quite close (except for ALAAM); and the Logit Model has the same recall for non-smoking behavior as the CRF Model.

	Non-Smoking			*Smoking*		
	Precision	Recall	F_1	Precision	Recall	F_1
Logit	0.771	**0.957**	0.854	0.688	0.250	0.367
LNAM	0.768	0.940	0.845	0.611	0.250	0.355
ALAAM	0.710	0.845	0.772	0.182	0.091	0.121
CRF	**0.776**	**0.957**	**0.857**	**0.706**	**0.273**	**0.394**

Table 4.5: Model predictions of next year's behavior.

Summing up, CRF Models appear to be capable of capturing social influence phenomena in networks. These models address both statistical and computational challenges of social influence in large networks. In particular, they incorporate the advantages of LNAM and ALAAM and correct the estimated coefficients and standard errors in Logit models (e.g. female effect in the comparison). They show the best performance in both goodness of fit and prediction. Furthermore, they have a fast and stable parameter estimation.

Now we will use the introduced models to study churn behavior in a virtual community.

4.2 Churn Behavior in Online Communities

The success of an online community is typically strongly related to the commitment and the active participation of its members. Thus, besides gaining new users, operators of online platforms also face the challenge to retain actual ones. So the question, why people leave a community becomes crucial and the analysis as well as the prediction of churn behavior gains more and more importance.

In this section we aim to identify different factors that drive quitting behavior. Apart from individual attributes of the users, we examine social network effects. In particular we want to find out whether quitting behavior is contagious and try to quantify the influence of quitting neighbors of a user in an online social network. Therefore, we apply the before-mentioned models.

Based on the literature (see Section 2.3.2) we phrase the following hypotheses:

- **Commitment hypothesis**
  - **Hypothesis 1 (H1):** Players are less likely to quit if they have a long time commitment in the game.
- **Achievement hypothesis**
  - **Hypothesis 2 (H2):** Players are less likely to quit if they achieve more in the game.
- **Social effects hypotheses**
  - **Hypothesis 3 (H3): (Community)** Players are less likely to quit if they are involved in an in-game organization.
  - **Hypothesis 4 (H4): (Quitting Together)** Players are more likely to quit if their partners quit.

4.2.1 Quitting in EverQuest II: Data and Descriptive Statistics

For this study we focused on a network of co-players in the Massive Multiplayer Online Role Playing Game (MMORPG) EverQuest II [Son16]. Virtual worlds and online games are widely used to study human behavior and social interaction. In EverQuest II each player controls a character and leads this character through different adventures. This includes the exploration of various environments in order to complete quests and to kill monsters. An important aspect of the game is that it fosters players' interactions. To solve different tasks the players can form groups or join guilds. If the tasks get harder this becomes a necessity.

We define churn behavior in the context of EverQuest II, if a player calls a certain number to cancel her subscription. At that time there was a gaming fee to pay every month and some players used to cancel their subscription several times in order to save that gaming fee but later joined the platform again. However, overall quitting is a very rare event in our data. Players are more likely to quit in the early stage of the game. Furthermore, solo players are three times more likely to display churn behavior than players with partner. In this discussion, we focus on the latter as we are interested in social influence effects. We considered two players as partners or co-players if they are solving some task together, e.g., killing a monster, at least twice, i.e., at two different days, to make

sure that there is a social relationship between the players rather than they only played together once coincidentally.

For our analysis, we use a data sample of one game server of EverQuest II. These data include statistics about players on that server for the months March to early September in 2006. However, as some of the models do not scale well, we restrict the analysis to one month, i.e., August 2006. We construct a cross-sectional network based on the activities in that month. Here, we consider all players who had a co-play relation during that time. The resulting network consists of 2587 nodes representing the players and 2320 links representing the co-play relations. Furthermore, 220 players out of the 2587 quit the platform either in August or September. Thus, only 8.5% of the individuals are quitters, i.e., exhibit churn behavior. In Figure 4.3 the partnership network is displayed. On average, a player has 1.8 co-play relations. However, most of the players only have one partner (i.e., 1553 players or 60.0%). One player, on the other hand, has 18 partners. Quitters and the links between them are displayed in red. When looking at the picture, it already seems as if there are a lot of players quitting together. However, we will test this impression statistically.

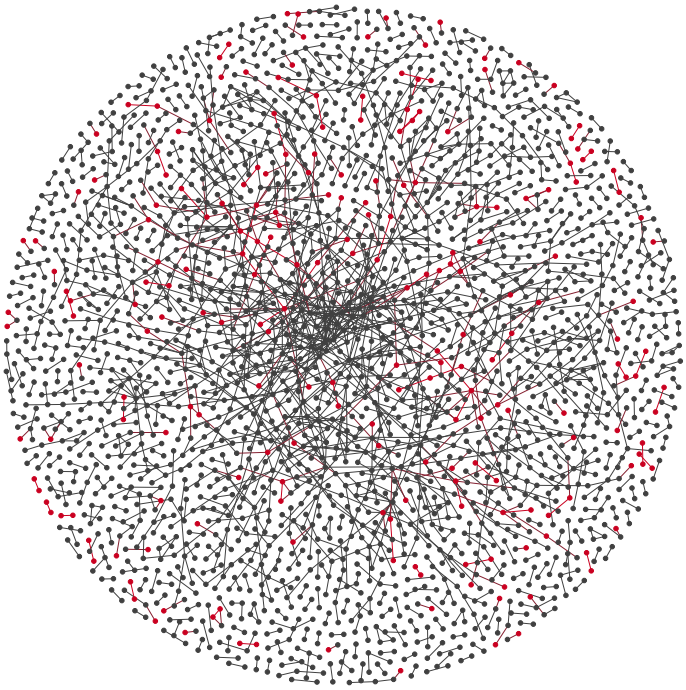

Figure 4.3: EverQuest II partnership network (220 quitters and links connecting them are displayed in red).

To test our hypotheses, we introduce the following measurements: Related to individual characteristics we consider the *character age*, i.e., the number of years since the player has created her character, which has been evolving since then, and the *number of rare*

items, i.e., all the rare items a player has obtained so far and which are particularity valuable. The first measurement we use to capture *commitment* and the latter to capture *achievements*. Related to network effects, we consider on the one hand *guild membership*, whether a player is part of an in-game guild, and the other hand *quitting together*, i.e., the social influence effect, which can be estimated by all models but the Logit Model. As a control variable we consider the *player's age*.

In Table 4.6 mean and standard deviation of the introduced measurements are listed and in Table 4.7 the correlations between them. All the measurements are negatively correlated with quitting behavior. However, the coefficients are rather small. Furthermore, all the measurements are weakly to moderately positively correlated among each other. The highest coefficient (i.e., 0.25) is between the age of the character and being member of an in-game guild. All correlations are highly significant.

Measures	Mean (SD)
Quitting	0.09 (0.28)
Character Age	1.12 (0.67)
Rare Items	134.47 (189.21)
In Guild	0.90 (0.30)
Player Age	33.15 (9.85)
$N = 2587$	

Table 4.6: Mean and standard deviation of the introduced measurements.

	Quitting	Character Age	Rare Items	In Guild	Player Age
Quitting	1.00				
Character Age	-0.11***	1.00			
Rare Items	-0.09***	0.16***	1.00		
In Guild	-0.11***	0.25***	0.12***	1.00	
Player Age	-0.10***	0.22***	0.18***	0.12***	1.00

Note: *p<0.05; **p<0.01; ***p<0.001; $N = 2587$.

Table 4.7: Correlation table of the introduced measurements.

4.2.2 Analysis and Results

In the same way as before we compare the four models regarding their ability to capture the setting. The results are displayed in Table 4.8.

Hypothesis 1 is supported by the Logit Model and the ALAAM. According to those models, players are in fact less likely to quit the game if they have a long time commitment. However, according to the LNAM and the CRF Model, this aspect has no significant impact. Hypothesis 2 is supported by all models, i.e., players are less likely to quit if they achieve more in the game. However, the coefficients are very small, in particular in the LNAM. Hypothesis 3 is supported by the Logit Model, the ALAAM and the CRF Model. For Logit and CRF the coefficient is almost the same, i.e., -0.578 for Logit and

		Dependent variable:			
		Quitting			
Hypotheses	Measures	Logit	LNAM	ALAAM	CRF
H1: Commitment	Character Age	-0.298**	-0.013	-0.290**	-0.209
		(0.111)	(0.008)	(0.107)	(0.112)
H2: Achievement	Rare Items	-0.002**	0.000**	-0.002**	-0.002**
		(0.001)	(0.000)	(0.001)	(0.001)
H3: Community	In Guild	-0.578**	0.031*	-0.392*	-0.579**
		(0.194)	(0.015)	(0.190)	(0.197)
H4: Quitting Together	Contagion		0.131***	2.276***	0.384***
			(0.008)	(0.227)	
Control	Player Age	-0.029***	0.002***	-0.025**	-0.026**
		(0.008)	(0.000)	(0.008)	(0.009)
	Constant	-0.425		-0.852**	-0.731***
		(0.282)		(0.311)	
	Activity			-0.260**	
				(0.099)	
	Partner Activity			-0.149*	
				(0.072)	
	N	2587	2587	2587	2587

Note: *p<0.05; **p<0.01; ***p<0.001

Table 4.8: Churn behavior: results of model comparison.

-0.579 for CRF, and clearly significant ($p < 0.01$). For ALAAM the coefficient equals -0.392 and it is weakly significant ($p < 0.05$). According to those models, players are less likely to quit if they are involved in an in-game organization. For the LNAM, on the other hand, the coefficient is weakly significant ($p < 0.05$) and positive (0.031). Thus, according to LNAM players are slightly more likely to quit if they are part of a guild.

Regarding hypothesis 4, quitting together is strongly supported by all models that are able to capture network effects. Furthermore, all coefficients are quite high compared to the other coefficients in the respective model but dissimilar. In the LNAM the contagion parameter equals 0.131, in the ALAAM it equal 2.276 and in the CRF it equals 0.384. Thus, players are more likely to quit if their partners quit as well.

Age is clearly significant and negative in the Logit Model, the ALAAM and the CRF Model, i.e.,the older a player the smaller the likelihood that she is quitting the game.

The coefficient is very similar these three models. For LNAM, on the other hand, the coefficient is significant and positive but very small. The constant term is negative and strongly significant for the ALAAM and the CRF Model. In the first case it equals -0.852 and in the latter -0.731. This strongly negative baselines confirm that usually they players do not display churn behavior, i.e., quitting is a rare event.

In the ALAAM again activity is included, which is significantly negative. It implies that players who are more active, i.e., who are well connected to other players regardless of their quitting behavior, are less likely to quit. Additionally, we also include partner activity, which is also significantly negatively and implies that players who have more active partners, i.e., partners with a lot of co-players that are quitter or not, are less likely to quit. Adding this parameter to the model made it more stable. However, within this setting, the parameter estimation with the ALAAM has several drawbacks. Because of the network size, i.e., 2587 nodes and 2320 edges, it took two days to estimate the model. Furthermore, the results are not stable. Compared to that, the parameter estimation with the LNAM took two hours and with the Logit and the CRF Models less than one minute.

4.2.3 Predictive Power of the Models

We now compare the Logit Model, the LNAM and the CRF Model regarding their ability to predict churn behavior. We do not consider ALAAM here as it is clearly geared towards statistical inference and performs badly in prediction tasks (see Section 4.1.3). Again we consider accuracy, precision, recall as well as the F_1 measure (see equations 4.5 to 4.8 at page 100).

However, this setting is also very challenging for the other models as quitting is a very rare event since only 8.5% of the users display this behavior. If the models Logit, LNAM and CRF are applied to predict a behavior, they assign a probability between 0 and 1 to each user that expresses the likelihood that she will adopt this behavior. Now, if the behavior is predicted typically a default cut-off of 0.5 is applied, i.e., if the probability is at least 50%, usually a prediction function would forecast that the person in fact adopts the behavior. In this setting we have to adapt the cut-off and choose a smaller one since the models are all strongly biased toward the non-quitting behavior. If we take the default cut-off of 0.5 the Logit Model and the LNAM predict quitting for non of the players; the CRF Model predicts this behavior for 10 players. However, those are all false-positive predictions.

If we, for example select a cut-off of 0.10 the Logit Model predicts 1761 (74.4%) non-quitters correctly and 108 (49.1%) quitters, leading to an overall accuracy of 72.7%. For the same cut-off, the LNAM predicts 1984 (83.8%) non-quitters correctly but only 26 (11.8%) quitters. However, due to the fact that the model almost always predicts non-quitting and obtains here a high hit rate, the overall accuracy is also comparatively high, namely 77.7%. The CRF Model predicts 1850 (78.2%) non-quitters correctly and

98 (35.5%) quitters, which leads to an overall accuracy of 75.3%, which is higher as the accuracy of the Logit Model.

Cut-Off: 0.10	*Non-Quitting*			*Quitting*		
	Precision	Recall	F_1	Precision	Recall	F_1
Logit	**0.940**	0.744	0.831	0.151	**0.491**	0.231
LNAM	0.911	**0.838**	**0.873**	0.064	0.118	0.083
CRF	0.938	0.782	0.853	**0.159**	0.445	**0.235**

Table 4.9: Predictive power of Logit Model, LNAM and CRF Model (using the cut-off 0.10).

In Table 4.9 precision, recall and the F_1 measure for both non-quitting and quitting behavior are listed. For the non-quitting behavior we see the effect that LNAM almost always predicts non-quitting. Thus, its performance with this regard is good but it is not capable of predicting quitting behavior. The highest precision for non-quitting behavior is achieved by the Logit Model. The CRF Model is second best for all precision, recall and F_1 measure with respect to non-quitting. When predicting quitting behavior, the CRF Model has the highest precision and F_1 measure and the Logit Model achieves the highest recall.

Now let's select a cut-off of 0.08. Here, the Logit Model predicts 1387 (59.0%) non-quitters correctly and 143 (65.0%) quitters, leading to an overall accuracy of 59.1%. The LNAM predicts 1495 (63.2%) non-quitters correctly and 69 (31.4%) quitters. Thus, the overall accuracy is 60.5%. The CRF Model predicts 1464 (61.9%) non-quitters correctly and 138 (62.7%) quitters, which leads to an overall accuracy of 61.9%, which is the highest accuracy.

In Table 4.10 the different measures are listed again. For the non-quitting behavior, the Logit and the CRF Model have the same precision that is higher than the precision of the LNAM. However, the LNAM has the higher recall followed by the CRF Model, which has overall the highest F_1 measure. Regarding the quitting behavior, again the CRF Model has the highest precision and F_1 measure and the Logit Model achieves the highest recall.

Cut-Off: 0.08	*Non-Quitting*			*Quitting*		
	Precision	Recall	F_1	Precision	Recall	F_1
Logit	**0.947**	0.586	0.724	0.127	**0.650**	0.213
LNAM	0.908	**0.632**	0.745	0.073	0.314	0.119
CRF	**0.947**	0.619	**0.748**	**0.133**	0.627	**0.219**

Table 4.10: Predictive power of Logit Model, LNAM and CRF Model (using the cut-off 0.08).

Overall, we see that it is very challenging to predict churn behavior in the given setting. the Logit Model and the CRF Model show a similar performance and this performance

is clearly better than the one of the LNAM. With respect to the F_1 measure, which combines precision and recall, the CRF Model appears to perform better.

There is also a high trade-off between the selection of the cut-off point and the false-positive results that occur, i.e., non-quitters that are classified as quitters. However, here in the discussion the focus is mainly to compare the different models and to examine their capability of capturing social influence rather than identify the best procedure to predict quitting behavior. This is why we also have not included predictor variables such as past quitting behavior. As we mentioned above, some players quit if they know they will play a lot in the following month to safe the gaming fee but will come back later when they have more time again. Our analysis has shown that this variable capturing past quitting behavior is the best predictor for future quitting behavior as well. However, some other effects disappear or have a smaller impact so we decided to exclude this variable.

Summing up, the results imply that the commitment of a player, her achievements and community effects decrease the likelihood that this player will quit the MMORPG EverQuest II. However, those results are not consistent with respect to the different models. On the other hand, we show that social influence cause a negative effect, quitting behavior might become contagious. This is clearly confirmed by all the three, LNAM, ALAAM and the CRF Model.

4.3 Sentiments in Online Travel Forums

Based on [NRW16], we aim to discuss an application setting for these type of models in this setting. We focus on user activities and interactions in the tourism domain. Here, in particular, the sentiments of the users are taken into consideration and we want to find out whether these sentiments are interrelated, i.e., we aim to answer the question: *Am I happy because my peers are happy?*

To examine this question, a travel related online forum is used where users are discussing their forthcoming trips. Social network analysis is applied to characterize the interactions between the users. To capture their emotions, a measure, which is constructed based on free-text comments in the forum, is assigned to the users. Here, text mining techniques and sentiment analysis are applied (see [NPW15, NRW16]).

Thus, in this section we have a continuous outcome variable capturing the behavior, i.e., the sentiments, rather than a binary one. Furthermore, we will assign weights to the network edges. Therefore, only the LNAM will be applied.

4.3.1 Analysis and Results

The analysis is done within a project with a start-up company. This company is an online marketplace where group tours to over 200 countries of the world can be compared, booked and discussed. Details about a tour including the points of interests that are visited, the length of the tour, etc. are provided by the respective tour operator. After the tour, a traveler can leave a tour review on the platform (see [NSSW14a]).

An important feature of the platform is the discussion within meets. In these meets users are given the opportunity to engage online with co-travellers before the tour starts. Typically tour related questions are discussed here. The messages are usually short and are often written in moments when users are excited, i.e., after booking a tour or before the departure. Meets are organized as threads, i.e., sequences of messages that are posted as replies to one another. Every user can start a meet and several meets related to one tour can exist. Meets provide the opportunity to study interactions and possible influence between users, thus they are the focus of the work presented here.

Data Sample

The data for the study were received from the company. Apart from the user generated free-text within the meets and reviews, the database contains meta-information about meets, tours and users of the platform. For each meet it is known when it was started, the comments it includes, when and by whom these comments were posted. Here the IP-addresses of the users are also stored. We use this IP-addresses to determine the country of the users as they typically participate in the discussions before a tour starts. It turned out that back in 2013 mainly Australians and people from other English speaking countries were using the platform. This clearly makes sense since the company was founded in Australia and only later moved to Europe.

Furthermore, it is known to which tour the meet is assigned to as well as the date when the respective tour started. However, the latter has to be indicated by the users themselves with no restrictions and is, thus, a bit noisy. The available information about tours encompasses a number of attributes including tour length, destination, tour operator, and maximum possible group size. User attributes include gender, location, birth date and language. Except for gender, these details are missing for the majority of the users. It is also known how active users are overall on the platform.

For the analysis, all meets that were posted on the platform within a 30 days period, i.e., April 2013, are analyzed. After dealing with some inconsistencies in the database and cleaning the data, the resulting sample has 3066 comments posted in 789 meets by 1270 distinct users. Thus, on average, each meet has 3.9 comments and each user posts 2.4 comments. Furthermore, 789 meets are related to 286 tours, i.e., per tour there are on average 2.8 meets taking place.

User Network

To study interactions and influence between the users on the platform, an undirected network is constructed the following way: the vertices of the network represent all users that were writing a comment in a meet in the selected period. Two users are connected by an edge if they were engaged in the same meet. Furthermore, a weight is assigned to the edge, which represents the number of different meets two connected users were part of. This process is illustrated in Figure 4.4.

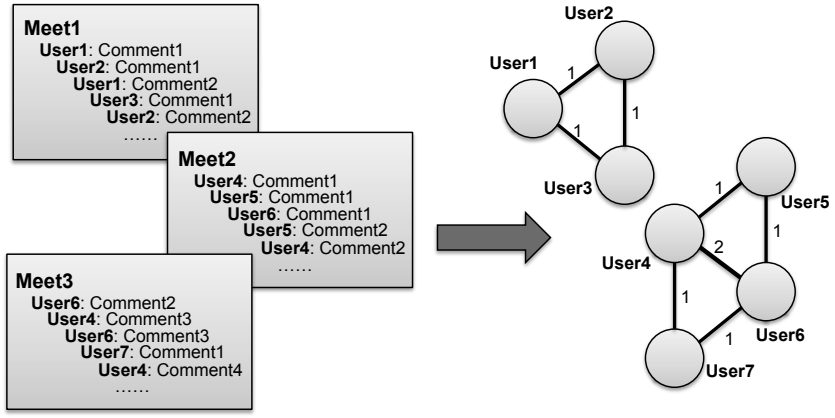

Figure 4.4: Construction of a user network based on the meets [NRW16].

In the resulting network, 1270 users are connected by 2055 edges. Thus, on average, each user interacted with 3.2 others. This is also the average degree in the network. The highest degree is 22, i.e., there is one user who interacted with 22 others. On the other hand, there are 345 isolates (27.2%) in the network. These users tried to initiate a conversation but nobody replied. Almost all edges have a weight equal to one; eleven edges have a weight equal to two, and one edge has a weight equal to three. This implies that only eleven pairs of users met in two different meets; and one pair of users even met in three different meets.

The network has a high number of small connected components. There are 228 connected components that consist of at least two nodes, and the largest component has 51 nodes. Thus, different regions of the network are hardly connected, but the nodes within a region are densely connected. This can clearly be seen in Figure 4.5, where the global structure of the network is displayed. This is not surprising and only reflects the semantics of the constructed network, namely, that each user is typically going only on one tour in a certain period and is, thus, participating only in those meets which are related to that specific tour.

In the sample of this work, the female/male ratio is almost 3/1: among 1270 users 941 are female and 329 are male. There is no significant difference between the average degree of male and female users in the network, although men have a slightly higher average degree (3.37 vs. 3.18).

Sentiment Scores

Focus of the discussion is the analysis of the emotions of the users and the interdependencies between those emotions. Thus, a measure, called sentiment score, is constructed with the aim to capture the state of mood of each user. This sentiment score is obtained with the help of a text mining procedure and is based on all free-text comments a user posted in April 2013.

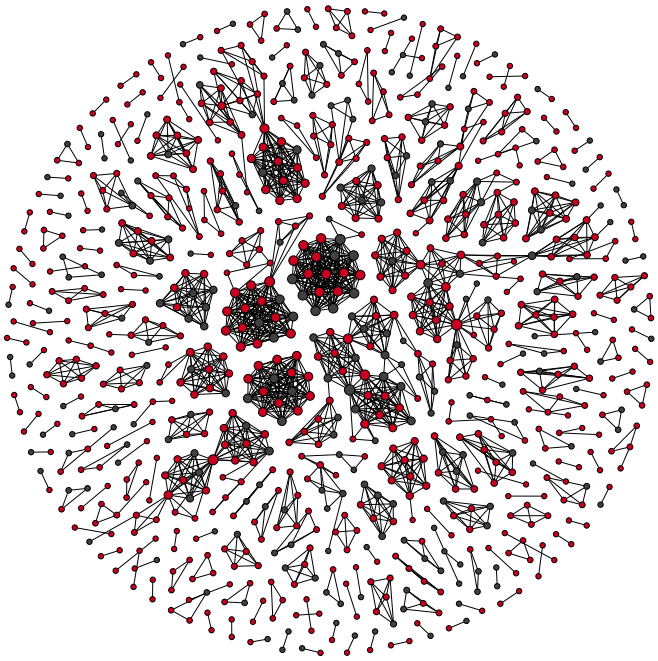

Figure 4.5: Global structure of the user network (without isolates; female users are displayed in red and male users in black) [NRW16].

The procedure is as follows. Firstly, tokenization and part-of-speech (POS) tagging of the comments are performed [BKL09]; afterwards, SentiWordNet [ES06, BES10] is applied. SentiWordNet assigns to each word both a positive as well as a negative score; where both of them can be zero (than the word is objective). However, note that in SentiWordNet a word with a specific meaning and POS tag is represented as a synset. Since a word can have different meanings depending on the context, a word can have several synsets, and all of them can have different positive and negative scores. For example, an adjective "poor" has three synsets. All of them have a positive score equal to 0, but the first one has a negative score 0, the second one has 0.125, and the last one has 0.5. To resolve this issue, the average of the scores of all synsets is used [TBT$^+$11]. The presence of negation in the text is addressed in the following way: Once a negation is encountered in the sentence, positive and negative scores for the rest of the tokens in the sentence are swapped [MSW$^+$11]. In this approach, emoticons are also taken into account. A sentiment score of 1.0 is assigned to positive emoticons and -1.0 to negative emoticons. Their values are not swapped after a negation. For each sentence the sentiment score is calculated as a difference between positive and negative scores per each word and then summed up. Such approach allows to accurately capture the overall sentiment in the sentence. For example, a sentence with an overall negative sentiment is "Sorry guys I've had to postpone my trip to Africa due to some unforeseen circumstances." whereas "Woo

can't wait :)" has an overall positive sentiment score. "How's everyone's packing lists going?", on the other hand is a rather neutral sentence.

Now, for each user her sentiment score is determined as an average of the scores of all sentences in all her comments posted in April 2013. The sentiment score of user 6 in Figure 4.4, e.g., is the average of the sentiment score of the sentences in her comment 1, comment 2 and comment 3. Regarding the overall distribution of the sentiment scores, the average sentiment score is 0.17, the median 0.13, the minimum sentiment score is -0.6 and the maximum sentiment score is 1.63. Thus, most of the sentiment scores are positive. When considering female and male users separately, it turns out that there is a significant difference between their average sentiment scores (0.19 vs. 0.11, $p < 0.001$). The reason for the positivity of the posted messages might be explained by the fact that future travelers are usually excited about their forthcoming tour.

Regarding the origin of users, the average sentiment scores of users from the US are significantly lower than those of users not from the US (0.10 vs. 0.17, $p < 0.001$); the same holds for Canadians (0.12 vs. 0.18, $p < 0.001$). On the other hand, the average sentiment scores of users from the UK are significantly higher than the average sentiment scores of users not from the UK (0.20 vs. 0.16, $p < 0.05$). For the other countries there are no significant differences.

Influence Models

To test whether the users influence each other regarding their emotions, Linear Network Autocorrelation Models are developed (see Equation 4.2 at page 94). The sentiment scores of the users are the outcome variable. The weighted matrix W represents the structure of the previously constructed network. This implies that only users can influence each other that are connected. As we have seen, there is a difference in sentiment scores for females and males. Thus, gender is included as a predictor variable. Furthermore, two dummy variables are constructed: the first indicates whether a user is from the US or Canada and the second whether a user is from the UK. Those dummy variables are included into the model as predictor variables since users from these countries have on average a significant smaller (and respectively larger) sentiment score compared to the other users. As control variables, the length of a tour (in weeks) and the number of comments written by a user are also included. Mean and standard deviation of the variables are displayed in Table 4.11. Furthermore, the correlations between the variables are shown in Table 4.12. As discussed previously, the sentiment scores are positively correlated to gender and origin UK and negatively correlated to origin USA or Canada. The length of a tour is weakly positively correlated to coming from the USA or Canada and moderately negatively correlated to coming from the UK; the latter is strongly significant. Of course a user is either from the UK or the English speaking part of North America. However, there is a significant negative correlation between those two variables as there is a number of users that are from somewhere else. Furthermore, length of a tour is weakly positively correlated to the number of comments a user is posting.

Variables	Mean (SD)
User Sentiment Scores	0.17 (0.21)
Gender Female	0.74 (0.44)
From USA or Canada	0.17 (0.37)
From UK	0.24 (0.43)
Length of Tour	2.8 (1.79)
Number of Comments	16.44 (27.81)
$N = 1270$	

Table 4.11: Mean and standard deviation of the variables.

	Sentiment Scores	Gender Female	From USA or Canada	From UK	Length of Tour	Number of Comments
Sentiment Scores	1.00					
Gender Female	0.17***	1.00				
From USA or Canada	-0.12***	0.02	1.00			
From UK	0.07*	-0.02	-0.26***	1.00		
Length of Tour	-0.04	0.05	0.08**	-0.29***	1.00	
Number of Comments	0.02	0.05	-0.03	0.01	0.09**	1.00

Note: *p<0.05; **p<0.01; ***p<0.001; $N = 1270$.

Table 4.12: Correlation table of the variables.

In Table 4.13 the results are displayed. We see the previously discussed effects: females and users from the UK are more likely to have a higher sentiment score. On the other hand, users from the US and Canada have typically a lower sentiment score. The model also shows that users who plan a longer trip are more likely to have a higher sentiment score. Furthermore, The results imply that a positive influence between the users' sentiment scores exists. Thus, the sentiment score of a user is influenced by the weighted linear combination of the sentiment scores of her peers; the more connections a user has, the higher the contribution of the network on her sentiment scores. Also, if two users meet in more than one discussion, the influence through this connection gets more important.

LNAM type of models do not scale well, thus the observation period is quite short. For robustness tests other observation periods have been used, and there are no significant differences in the results. We also fit a model without taking into account the network structure, which is equivalent to OLS. The results are almost the same apart of course from the contagion effect that cannot be captured by the OLS.

Summing up, the results imply that the answer to the question stated at the beginning of the seciton, i.e., *Am I happy because my peers are happy?*, is yes. The emotions of the users appear interdependent; a user seems to be influenced by the emotions of all her network connections. Thus, the results imply that in the context of tourism positive emotions can be seen as an asset that influences others. However, the same is true for negative emotions. An obvious question would be, if bad mood in a forum can be changed

	Dependent variable:
	Sentiment Scores
Female	0.140***
	(0.011)
From USA or Canada	-0.035*
	(0.016)
From UK	0.058***
	(0.013)
Length of Tour in Weeks	0.011**
	(0.003)
Number of Comments by User	0.000
	(0.000)
Contagion	0.019**
	(0.007)
Pseudo R^2	11.0%
N	1270

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4.13: Result of the LNAM.

by positive influence. However, such a study might raise some ethical problems as in the case of the "Facebook Study" [CPD15]. Another issue is how sentiments in discussions before the tour influence the formation of the destination image and affect the overall satisfaction from the travel experience. This would enhance the study of destination branding and image [KD15].

However there are also some limitations. The assumption in this study is that all users in a thread are interacting with each other, i.e., their interactions are represented by an undirected network. This assumption is reasonable because of the short observation period. However, in this analysis it is not taken into account how many messages are posted within one thread. In a next step this will be taken into consideration when constructing the weighted network as more interactions might reinforce the influence. The sentiment scores are extracted and assigned using an automated procedure. Although this approach has its limitations, it is state-of-the-art and well-accepted. One crucial aspect is here the issue of social influence vs. social selection (see also 2.2.3). This model shows that the there is a contagion effect of the sentiment scores. However, part of the effect might be due to social selection. Maybe users who are in a good mood rather participate in conversations where positive sentiments are already prevalent. What we observe might be a consequence of both. Here, for sure, further analysis is necessary.

4.4 Discussion

In this chapter we discuss exiting approaches that are able to account for interdependencies between nodes when modeling an outcome behavior. This is very challenging, in particular, as we are interested in cross-sectional data. ALAAMs tackle this problem by a costly simulation process. This approach works well for small networks but does not scale. Furthermore, it is geared toward statistical inference and not prediction. LNAMs are another type of network models that are able to account for interdependencies between nodes. However, social influence is modeled as a weighted linear combination of the neighbors' behaviors, which might be too simplifying in many settings. Furthermore, they require a continuous outcome variable. Another important issue is that the estimation process takes quite long for bigger networks. Thus, it might be tedious in some situations to apply network models, in particular, as logistic regression performs in practice very fast and apparently quite accurate. However, there are certain issues when applying regression models to network data, as they might lead to wrong estimations of the standard errors [FCS10].

We propose CRF Models as an alternative. In this chapter we show that they provide a good alternative to both logistic regression models as well as the network models LNAMs and ALAAMS. CRF models address both statistical and computational challenges of social influence in large networks. In the discussed empirical examples they overall performed best, particularly when predicting the "new" behavior. Furthermore, they show fast and stable parameter estimations.

However, in both settings, i.e., smoking behavior and churn behavior, the models lead to slightly different results regarding size and significance of the different coefficients as well as the influence parameters. Although we can conclude, which results are more reasonable and assess their accuracy, we do not know the ground truth. Thus, what will be done next is to compare the different models in a systematic way. We will simulate data where we specify the strength of the influence mechanisms; thus, we can design the ground truth. This will also help to distinguish between social influence and social selection mechanisms, which is particularly an issue in cross-sectional settings. We tried to address it here with the help of longitudinal information. For example, in the case of the friendship network we use the network structure at time t_1 and the behavior at time t_2 to ensure that the behavior is a consequence of the network and not the other way round. However, it might be an issue in the setting of the sentiment scores. Here, a deeper analysis is required.

CHAPTER 5

Social Influence and Performance: A Multi-Level Analysis

In this chapter we aim to discuss complex settings where all levels of information interact (see part A, B and C in Figure 5.1). Utilizing one more time the example of smoking teenagers, we illustrate how information at the group level can be included in network models. Thus, we combine insights from the individual level, the group level and the network level.

Next we conduct a study that focuses on team performance and aims to model duration and outcome of head-to-head competitions, where the team's overall goal is to beat the opponent. We use this setting to observe multilevel factors that influence the relative performance of the teams. Those factors include compositional factors capturing the individual level, relational factors capturing the network level and ecosystem factors capturing the group level.

5.1 Social Influence at Three Levels of Information

In Chapter 3 we discuss the setting of smoking teenagers at the group level, demonstrating how this extends the individual level. We identify four groups or clusters that show clear differences regarding the smoking behavior of the pupils between the respective clusters. In particular, cluster 1 comprises a relatively high number of smokers, 43.4% of all members of this cluster smoke, whereas only 6.9% of members of cluster 4 are smokers (both at time t_2). This information is included in logistic regression models with smoking behavior as an outcome variable. In this way, we are able to compare it other individual attributes, i.e., gender, being in a romantic relation, having a smoking sibling and drinking behavior. We show that the model improves considerably when information about cluster membership is included (see Table 3.10).

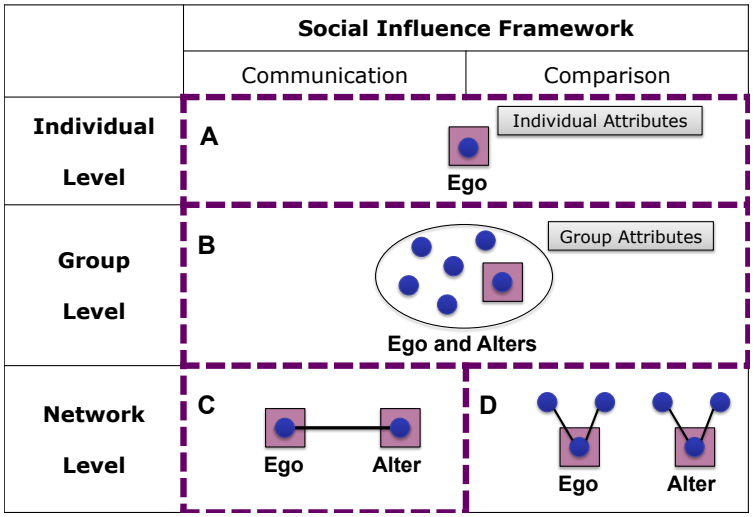

Figure 5.1: Social influence processes and levels of information.

In Chapter 4 we discuss smoking behavior among teenagers at the network level. Friendship relations between the pupils are considered when analyzing their smoking habits. Social influence models for networks are applied to this setting. These models detect a contagion effect, which implies that the smoking behavior of a pupil is influenced by the smoking behavior of friends (see Table 4.3).

In a next step we combine these two levels by integrating the cluster membership as a nodal attribute in the CRF Model. The results are displayed in Table 5.1. We list the results of the basic Logit Model (i.e., Logit M1), of the Logit Model including cluster membership as predictor (i.e., Logit M2), of the CRF Model (i.e., CRF M1) and of the CRF Model including the cluster membership as nodal attribute (i.e., CRF M2). We see that in both cases, the Logit Model and the CRF Model, the intercept term is not significant any more as soon as cluster membership is included. Gender is not significant in any of the models but in the model Logit M1. Being in a romantic relationship is marginally significant in all the four models. A smoking sibling is marginally significant in all models but in CRF M1, where it is clearly significant. Drinking behavior is highly significant in all the four models. The information about cluster membership is significant in both models, Logit M2 and CRF M2. In the latter it is slightly less significant. Furthermore, the contagion effect in CRF M2 does not considerably change compared to CRF M1 if the cluster membership is included.

With respect to the predictive power of the models it turns out that when information about cluster membership is included the Model Logit M2 performs better than the Model Logit M1. However, there are no changes in the case of the CRF Models. With respect to the smoking behavior at time t_2, Logit M1 predicts non-smoking correctly in 111 cases and smoking in 11 cases; Logit M2 predicts non-smoking correctly in 120 cases and smoking in 12 cases. Regarding the CRF Models, both predict non-smoking

	Dependent variable:			
	Smoking			
	Logit M1	Logit M2	CRF M1	CRF M2
Constant	-2.568***	-0.497	-3.033***	-1.042
	(0.396)	(0.782)	(0.398)	(0.795)
Female	0.719*	-0.081	0.331	-0.391
	(0.436)	(0.532)	(0.443)	(0.545)
Relation	0.847*	0.836*	0.894*	0.914*
	(0.449)	(0.472)	(0.459)	(0.482)
Sibling smokes	0.892*	1.021*	1.007**	1.056*
	(0.502)	(0.533)	(0.510)	(0.548)
Drinking	1.418***	1.271***	1.320***	1.182***
	(0.423)	(0.438)	(0.434)	(0.449)
Cluster		-0.834***		-0.771**
		(0.305)		(0.307)
Contagion			0.448***	0.412***
N	160	160	160	160

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5.1: Smoking behavior: results of models that integrate different levels of information.

correctly in 119 cases and smoking in 13 cases. Thus, the additional information about cluster membership does not improve it. Regarding precision and recall, we calculate the F_1 measure as in Section 4.1.3. For non-smoking behavior it is almost the same for Logit M2, on the one hand, and the CRF Models, on the other hand, namely 0.896 vs. 0.895. Regarding the prediction of smoking behavior, Logit M2 performs still worse than the CRF Models, namely 0.462 vs. 0481. The basic logic model, i.e., Logit M1, performs worse with a F_1 measure of 0.883 for non-smoking prediction and of 0.415 for smoking prediction (see Table 4.4). The same patterns can be observed for the smoking behavior at time t_3.

Incorporating network structure as covariates into regression models is in general not recommended as independence assumption can be violated; this, in turn, might lead to mis-specified models [Sni11]. For example, in our setting, it would be tempting to include the number of smoking friends into the logistic regression model. However, as this variable refers to other pupils that are as well in the data sample, it causes a problem. The attributes of the pupils are not independent any longer, but this is required. A

different situation is presented in Section 5.2, where we use the density of within team networks in regression models. Here, the teams are separated entities with independent internal structures.

5.2 Social Influence and Performance in Multi-Team Systems

To understand, which factors impact the performance of a team is a complex endeavor. Typically, it will not be a single factor but combinations of factors at different levels that have positive or negative influence on the success of a team. The setting that we aim to study here is even more complex. We focus on relative performance in team-vs-team competitions in the MOBA game Dota 2. In this game two teams of five members are competing against each other and the only goal is to defeat the opposing team. Thus, a team has to constantly react to the opponent's activities. The matches in Dota 2 do not have a fixed length. Thus, we examine two dimensions of winning behavior, i.e., winning or losing, on the on hand, and the duration of a match, on the other hand. This is illustrated in Table 5.2. In general, a team that wins quickly, i.e., in a short duration, is clearly better than its opponent. On the other hand, if the match is already taking long there might not be much difference in the performance of the two teams and its almost a tie situation. In the end random effects might cause winning or losing.

Results \Duration	Short	Medium	Long
Win	Dominating	Good team	Marginally better
Loss	Dominated	Bad team	Bad luck

Table 5.2: Two dimensions of outcome: results vs. duration

5.2.1 Performance Factors and Duration

Our aim is to identify factors at different levels that influence the performance of a virtual team in these direct competitions. Based on the literature (see Section 2.3.3), we focus on three levels of factors in this context. These categories are related to the composition of a team, to relations within a team and to relations within the ecosystem of teams.

Compositional factors represent a collection of attributes of the team members such as individual skills and expertise and lay the foundation for achieving team goals. *Relational factors* represent the social bonding among team members and facilitate a better collaboration among them. *Team ecosystem factors* measure inter-team relations and describe the external environment of teams. As discussed in Section 2.3.3, compositional and relational factors, are clearly defined and well characterized in literature. Additionally, we propose team ecosystem factors as a separate category to capture the impact of inter-team relations on the performance of a team in more detail. If team members have played

in many teams with different combination of other players they had the opportunity to learn different approaches and tricks. Based on this we propose the following hypotheses:

- **Compositional Factors**
  - **Hypothesis 1 (H1):** Teams with higher players' skills are more likely to win.
- **Relational Factors**
  - **Hypothesis 2 (H2):** Teams with players with more previous collaboration ties are more likely to win.
- **Team Ecosystem Factors**
  - **Hypothesis 3 (H3):** Teams with players who played in many different teams are more likely to win.

With respect to the introduced Social Influence Network (see Table 1.1) the compositional factors capture the individual level, the relational factors capture the network level and the team ecosystem factors capture the group level. Although the latter is related to the idea of inter-team connections, we assign it to the group level as it is here not explicitly modeled by network structure and rather represents the global context. In this sense, ecosystem factors determine the position of a team within the complex team ecosystem.

Dota 2 matches do not have a specific length, which provides the unique opportunity to explore if and how these three types of factors influence the duration of a match.

When applying the traditional input-process-output model of team performance, the three categories of factors outlined above refer to the "inputs" to this model. They characterize the essential team attributes and relation patterns. However, to understand how these attributes and patterns actually interact and influence team performance we need to understand the dynamics of the "process" as this phase directly affects the "output".

In previous work, when studying teaming and how different factors influence the outcome, the detailed process of team collaboration has typically depended on the nature of the tasks. We, on the other hand, are focusing on a team-vs-team setting to separate the different processes. In a team-vs-team competition, the overall objective is to defeat the opponent and there is typically no pre-specified task.

The duration of a competition provides an approximate measure of complex interactions during the whole teaming process. We will use this duration to isolate the basic mechanisms of compositional, relational and team ecosystem factors in team collaboration. Therefore, we propose the following research question:

- **Team Process and Duration**
  - **Research Question: (RQ):** How do different performance factors change the amount of time a team takes to win?

5.2.2 Methods

Dota 2 Game Setting

Dota 2 is a MOBA Game produced by Valve [Cor14], where two teams, named the Radiant and the Dire, compete against each other. Each of the teams consists of five players and they are located at the opposite corners of the gaming map: Team Radiant at its lower left and team Dire at its upper right (see Figure 5.2). To win a match, a team has to destroy the opponents' Ancient, i.e., a massive structure within a team's stronghold that is guarded by two towers. Although the Radiant side and the Dire side are conceptually the same, there are a number of design differences between them. The environment of team Radiant, for example, is brighter and friendlier than the dark and gloomy environment of team Dire.

Each player controls a character called hero. These characters evolve during a match. They acquire experience, which helps them to level up, as well as gold, which can be used for buying items. There exist more than 100 hero characters in Dota 2, each of them with different attributes and abilities and different ways to evolve. This opens up many possibilities and makes the game very complex.

Heroes can die, but revive after a certain period. The length of this respawn time in seconds is computed by 4 $\times$ hero level but can be decreased with gold. Each match starts from scratch and takes on average about 45 minutes. However, there is no fixed length.

In order to ensure a fair match, Dota 2 utilizes a matchmaking system (1) to assign players to a team and (2) to match the opposing teams. Although the detailed algorithm has not been disclosed by Valve, it is known that the matchmaking mechanism strives to match players of similar skills and experiences against each other. The experience of a player is defined by the number of matches the player has played before and the skill measure is related to the performance of the player in those previous matches. However, also other hidden variables are taken into account when assigning the players into the opposing teams [Inc14].

Dota 2 was officially released in July 2013, but before that it had been available as a beta version with limited access since 2011. Dota 2 is a very popular and high-paying e-sports game; already in its beta phase, professional Dota 2 tournaments were taking place.

Data Samples and Measurements

Based on a game log of all Dota 2 matches in year 2011, we select 64,643 sample matches that were played in the second week of December 2011 (December 8^{th} to 14^{th}). In these matches, moreover, no hero is computer-controlled, each team consists of five human players. Since one central goal of our analysis is to study the impact of different factors on the performance of teams, we want to make sure that the sample matches are completed with a clear winner. As described previously, a match is completed if one team destroys the Ancient of the other team. The two towers guarding this Ancient have to be demolished before the Ancient can be attacked. Thus, we exclude matches

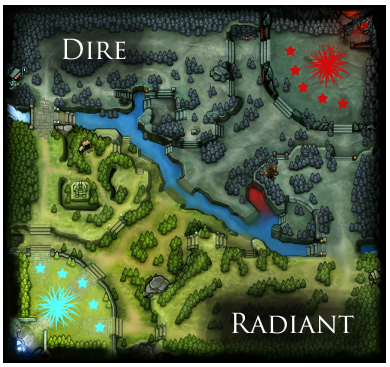

Figure 5.2: Dota 2 Gaming Environment. Note: Team Radiant is located at the bottom-left; team Dire at the top-right [PNCM+13].

where one team wins because the other team abandons the game. With these criteria, we obtain 62,034 matches with a clear winning/losing situation.

For all players and teams in the 62,034 sample matches, we use their activities before December 8^{th} to construct their game statistics and measure their skills, relations and team interactions.

The performance of a team is related to the skills of its members. In the complex and competitive setting of Dota 2 it is important for a team to have the abilities to attack, to defend, and to apply certain strategies to win a match. We relate these abilities to the in-game statistics in the following way: The number of enemy heroes a player kills in a match represents his or her attacking skill. Having a high number of kills is relevant since a player gains experience and gold from kills, which increases the chance of winning. The number of times a player's hero gets killed in a match captures his or her defending skill. A player usually strives for having a low number of deaths since the player loses a certain amount of gold and has to wait some time to revive. In Dota 2, denying is regarded as a complex cooperation strategy. A player gets a deny point if she kills an allied unit before the opponent is able to do so. The deny strategy prevents the enemies from gaining experience and gold.

To capture the skill statistics for each team, we first calculate individual player statistics based on the matches in the previous week and then aggregate at the team level. For example, the individual player's kills statistics is the mean of the numbers of kills in her previous matches. For each team, the team kills statistics (abbreviated as *"team kills"*) is the average of the individual kills statistics of all five team members and measures the overall attack skill of a team. To capture the defense skill, we compute individual player's death rate as the mean of the death-to-kill ratios in her previous matches. For each team, the team death rate (abbreviated as *"death rate"*) is the average of the individual death rates of all five team members and measures the overall defense skill of a team. Similarly, the team deny rate (abbreviated as *"deny rate"*) is the average of the individual

deny-to-kill ratios of all five team members and measures the team's ability to apply complex deny strategies.

The number of previous co-play relations (abbreviated as *"co-play"*) measures the previous relations among players in a team. A co-play relation between two players is given if they have played together at least twice in all previous matches. The range of the co-play measure is from 0 to 10 in a team of five and more co-play relations imply that the team members are more familiar with each other.

The number of unique partners (abbreviated as *"partners"*) measures the overall experience of playing in other teams for all members in a team. For each team we compute the total number of unique co-play partners (other than the current teammates) the five members have played together with in all previous matches; i.e., all unique players that are not in the team to whom the five members have co-play relations.

Note that the skill measures are constructed based on the matches of the previous week to get a more accurate estimation of a player's performance, whereas the co-play and partners measurements are constructed based on all previous matches to detect their potential interactions in the past.

5.2.3 Modeling Relative Outcome

Match Categories

To capture the impact of the compositional, relational and team ecosystem factors on the outcome of a match in this team-vs-team setting, we develop models of relative performance. Therefore, we take the perspective of one team. Thus, the dependent variable shows whether team Radiant beats team Dire in a match. The independent variables capture relative advantages by taking the differences (Δ) in team measurements, i.e., team Radiant minus team Dire. To model the relative outcome, we use binary logistic regression.

One main objective of our analysis is to study the impact of the different factors with regard to the duration of a match. Thus, as a first step we introduce different duration categories to find out whether or not we can find associations between the impacts of the factors and the length of a match. With respect to Table 5.2, we focus on the different columns whereas the results, i.e., winning or losing, represent the dependent variable. However, we introduce five duration categories rather than three applying the following procedure: The distribution of the durations of the 62,034 matches shows that on average a match lasts 2,791 seconds with a standard deviation (SD) of 668 seconds. We define five categories (see Figure 5.3): short, medium low, medium, medium high, and long duration matches using cut-off points 1455, 2123, 3459, and 4127 seconds (i.e., two SD below the mean, one SD below the mean, one SD above the mean, and two SD above the mean respectively).

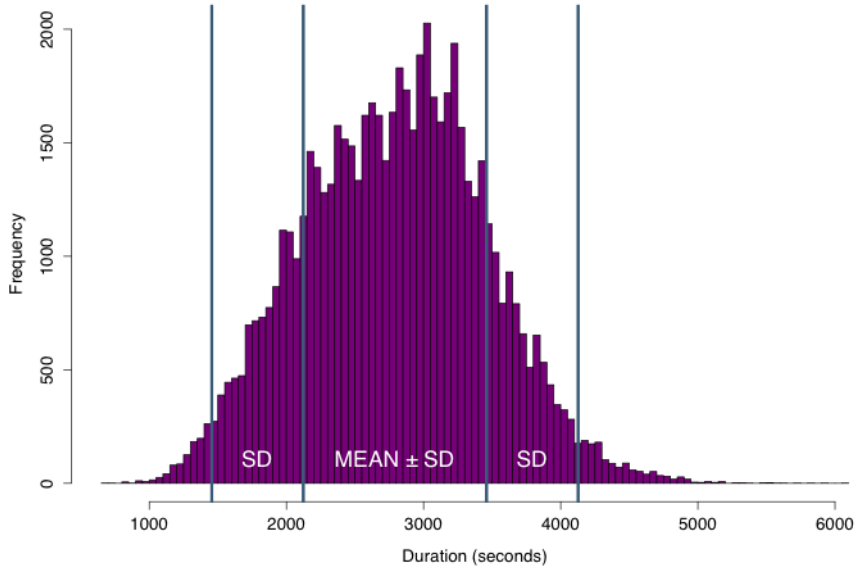

Figure 5.3: Five sample categories with different match durations.

Descriptive Statistics

The correlations between the variables for all matches ($N = 62,034$) are listed in Table 5.3. We see that the relative skill measures of a team are all weakly correlated with Radiant wins. Teams that have more kills than their opponents also tend to have a higher death rate and a higher deny rate respectively. Furthermore, the number of co-play relations is weakly correlated with the skill measures.

When studying the correlations for the single match categories, the results are quite similar (not shown). This is particularly true for medium low matches and for medium matches. For longer matches, the correlations between Radiant wins and the other measures are not significant. Furthermore, the correlation between Δ co-play and Δ partners is not significant for short and long duration matches.

	Radiant wins	Δ Team kills	Δ Death rate	Δ Deny rate	Δ Co-play	Δ Partners
Radiant wins	1.00					
Δ Team kills	0.02***	1.00				
Δ Death rate	-0.02***	0.31***	1.00			
Δ Deny rate	0.04***	0.20***	0.43***	1.00		
Δ Co-play	0.03***	0.26***	0.21***	0.15***	1.00	
Δ Partners	0.05***	0.13***	0.13***	0.13***	-0.10***	1.00

Note: $N = 62,034$; *p<0.05; **p<0.01; ***p<0.001

Table 5.3: Correlations of variables in relative outcome models.

Table 5.4 shows the descriptive statistics including means and standard deviations for all team matches as well as the breakdowns of Radiant/Dire teams and winning/losing teams. The average team kills are slightly higher for Radiant and for winning teams than for Dire and for losing teams. For winning teams the mean of co-play is also slightly higher than for losing teams. Dire has more co-play relations on average than Radiant. Radiant and winning teams have on average a higher number of unique partners than Dire and losing teams.

	All Teams	Team Radiant	Team Dire	Winning Teams	Losing Teams
Team kills	4.78 (2.52)	4.81 (2.51)	4.75 (2.54)	4.81 (2.52)	4.76 (2.52)
Death rate	1.27 (0.73)	1.28 (0.72)	1.27 (0.73)	1.26 (0.72)	1.28 (0.73)
Deny rate	0.81 (0.63)	0.81 (0.62)	0.81 (0.63)	0.82 (0.63)	0.8 (0.62)
Co-play	1.32 (1.81)	1.09 (1.50)	1.56 (2.04)	1.36 (1.84)	1.29 (1.77)
Partners	12.31 (16.24)	12.74 (16.74)	11.88 (15.72)	12.63 (16.75)	11.98 (15.71)
N	124,068	62,034	62,034	62,034	62,034

Table 5.4: Means and standard deviations of team statistics.

The descriptive statistics for the differences in team measures by match category are listed in Table 5.5. In short and medium low duration matches, team Radiant wins more often than Dire. In medium duration matches it is balanced; and for longer durations Dire predominately wins. Thus the longer the match duration, the fewer matches are won by Radiant. Although it is not known what causes this effect, this early game Radiant advantage is well known by the community [Liq14].

Especially in short duration matches, Radiant has a higher average team kills than Dire: in every 3.7 matches, Radiant has one team kills more than Dire. For the other duration categories, this measure is more balanced. For Δ death rate and Δ deny rate there are no big discrepancies between the duration categories. The number of average co-play relations is higher for Dire for all durations; however, the longer the duration the bigger the differences. In short duration matches, Dire has on average one more co-play than Radiant in one out of three matches. In long duration matches, on the other hand this is the case in one out of two matches. Radiant has more average partners in all categories. Here, the differences vary even stronger. For short duration matches, Radiant has on average one more unique partner than Dire in every 0.6 match. For long duration matches it is only one more unique partner every 1.6 matches.

Models and Results

Table 5.6 shows three models for short duration matches with different sets of measures in order to illustrate the prediction power of compositional, relational and team ecosystem factors. The first model includes compositional measures and studies the impact of player skills on the team outcome. Here, Δ death rate and Δ deny rate are significant. Teams with a higher death rate than their opponents are less likely to win. A difference of one

	Short	Medium Low	Medium	Medium High	Long
Radiant wins	0.56	0.53	0.50	0.47	0.46
Δ Team kills	0.27 (2.38)	0.07 (2.25)	0.06 (2.30)	0.02 (2.39)	-0.03 (2.45)
Δ Death rate	0.03 (0.84)	0.01 (0.82)	0.01 (0.84)	0.01 (0.85)	0.01 (0.87)
Δ Deny rate	0.00 (0.61)	0.01 (0.63)	0.00 (0.63)	-0.01 (0.64)	0.01 (0.64)
Δ Co-play	-0.33 (2.2)	-0.44 (2.25)	-0.48 (2.2)	-0.46 (2.25)	-0.49 (2.25)
Δ Partners	1.68 (16.69)	1.41 (16.05)	0.77 (13.38)	0.55 (11.65)	0.64 (10.77)
N	1,078	9,489	41,830	8,285	1,352

Table 5.5: Means and standard deviations of differences in team measures by duration category.

in the team death rate leads to a 70% odds ratio to win in a short duration match. On the other hand, teams with a higher deny rate than their opponents, are more likely to win. Here, a difference of one in the team deny rate leads to a 168% odds ratio to win.

The second model in Table 5.6 contains in addition the relational measure co-play. This measure is significant; teams with more co-play relations than their opponents are more likely to win. One more co-play leads to a 109% odds ratio to win. Both Δ death rate and Δ deny rate are still significant; and the coefficients of these measures are almost the same as in the first model.

The third model comprises all three types of measures, i.e., compositional, relational and ecosystem. We see that Δ partners is significant. Teams with more outside partners than their opponents are more likely to win in a short duration match. One more partner in one team leads to a 102% odds ratio to win. However, also Δ death rate, Δ deny rate and Δ co-play are still significant; and the coefficients of these measures do not change a lot. For all three models the intercept term is positive and significant whereas Δ team kills is not significant in any of these models. The variance that is explained increases from 4% in the first model to 5.1% in the second and 7.8% in the third model.

In summary, to win quickly (i.e., in a short duration match), it is an advantage for a team to consist of players who have better defense skills than their opponents, are able to apply complex strategies, have co-play experience within the team and a higher number of outside partners. In terms of variance explained, relational and the team ecosystem measures are as important as the compositional measures.

For all the other duration categories we applied the same stepwise procedure and compared the three types of models (here, we only list the results of the third models comprising compositional, relational and team ecosystem measurements – see Table 5.7).

The variable Δ team kills is not significant in most of the cases. Only in medium duration matches a small advantage can be observed; one more average team kills leads to a 101% odds ratio to win. The measure Δ death rate has a significant negative impact for medium duration matches and below. The shorter the duration, moreover, the higher this impact: a difference of one in the average death rate leads to a 61%, 73% and 88% odds

		Dependent variable:		
		Radiant wins		
Hypotheses	Measures	Compositional	Compositional + Relations	Compositional + Relations + Ecosystem
H1: Skills (attack)	Δ Team kills	0.01	0.00	-0.02
		(0.03)	(0.03)	(0.03)
H1: Skills (defense)	Δ Death rate	-0.44***	-0.47***	-0.50***
		(0.09)	(0.09)	(0.09)
H1: Skills (strategy)	Δ Deny rate	0.52***	0.51***	0.47***
		(0.12)	(0.12)	(0.12)
H2: Relations	Δ Co-play		0.09**	0.10***
			(0.03)	(0.03)
H3: Team	Δ Partners			0.02***
				(0.00)
	Intercept	0.26***	0.29***	0.28***
		(0.06)	(0.06)	(0.06)
	Pseudo R^2	0.040	0.051	0.078
	N	1,078	1,078	1,078

Note: *p<0.05; **p<0.01; ***p<0.001

Table 5.6: Relative outcome models for short duration matches.

ratio advantage to win in short, medium low and medium duration matches respectively. The measure Δ deny rate has a positive impact for all but long duration matches. Also here, the impact is strongest for short durations and decreases with duration category: a difference of one in the average team deny rate leads to a 160%, 138%, 117% and 112% odds ratio to win in short, medium low, medium and medium high duration respectively.

The measure Δ co-play has a significant positive impact in medium and shorter duration matches: One more co-play relation than the opponent leads to a 111% (short duration), 109% (medium low duration) and 102% (medium duration) odds ratio to win. Further, the measure Δ partners has a significant positive impact in short, medium low and medium duration matches; one more unique partner than the opponent leads to a 102% (short duration), 101% (medium low duration) and 101% (medium duration) odds ratio to win. The intercept term is significant for short, medium low and medium high duration matches; in the first two cases it is positive and in the latter it is negative. The amount of variance that is explained by the models is very modest: 7.8% for short duration, 4.4%

for medium low duration, 0.7% for medium duration, 0.3% for medium high duration and 0.5% for long duration matches.

Clearly the outcome is harder to predict the longer the duration of the match. Especially for medium high and long duration matches the outcome is highly unpredictable. As the duration gets longer, it appears that random effects (or at least effects not accounted for in our model) determine who will win.

		Dependent variable:				
		Radiant wins				
Hypotheses	Measures	Short	Medium Low	Medium	Medium High	Long
H1: Skills (attack)	Δ Team kills	-0.02 (0.03)	0.02 (0.01)	0.01** (0.00)	-0.01 (0.01)	0.00 (0.02)
H1: Skills (defense)	Δ Death rate	-0.50*** (0.09)	-0.31*** (0.03)	-0.13*** (0.01)	-0.07 (0.03)	-0.07 (0.07)
H1: Skills (strategy)	Δ Deny rate	0.47*** (0.12)	0.32*** (0.04)	0.16*** (0.02)	0.11** (0.04)	-0.03 (0.10)
H2: Relations	Δ Co-play	0.10*** (0.03)	0.09*** (0.01)	0.02*** (0.00)	0.02 (0.01)	0.06 (0.03)
H3: Team	Δ Partners	0.02*** (0.00)	0.01*** (0.00)	0.01*** (0.00)	0.00 (0.00)	0.00 (0.01)
	Intercept	0.28*** (0.06)	0.16*** (0.02)	0.00 (0.01)	-0.12*** (0.02)	-0.13 (0.06)
	Pseudo R^2	0.078	0.044	0.007	0.003	0.005
	N	1,078	9,489	41,830	8,285	1,352

Note: *p<0.05; **p<0.01; ***p<0.001

Table 5.7: Relative outcome models for all durations.

A stepwise regression for medium low duration matches shows that the explained variance increases from 2% for a model that contains only compositional measures, to 2.8% for a model with compositional and relational measures to 4.4% for the full model which includes also ecosystem effects. As in the case of short duration matches, the two sides – compositional measures on the one side and relations plus team ecosystem measures on the other side – have similar contributions. For medium duration models the explained variance equals 0.5% for a model with only compositional measures and does not increase when adding the co-play measure. However, in the full model it increases to 0.7%.

Summing up, hypotheses H1 (defense and strategy), H2 and H3 are confirmed for short duration matches, medium low duration matches and medium duration matches. In order

to win, a team should have the skills to protect itself and to apply complex strategies. Furthermore, it is a clear advantage if the members share prior co-playing experiences and if they have played in the past in other teams with more unique other partners. Hypothesis H1 (attack) is only significant in medium duration matches.

Relations and team ecosystem have a slightly higher impact in short duration than longer duration matches. That is, they are more likely to help teams win more quickly. However, if a team does not win quickly, the skills differential between the teams gains increasing importance compared to relations and team ecosystem. This is consistent with (Balkundi and Harrison, 2006) since results of their meta-analysis showed that the impact of relations on performance decreases as the team members get familiar with each other and their tasks.

For matches that last very long, none of the factors are significant and predicting the outcome is not possible.

5.2.4 Modeling Duration

Performance and Duration

In the previous section we show that compositional, relational and team ecosystem factors influence the outcome of a match in each duration category in a different way. Now we want to establish the association between the factors and the duration of a match more explicitly to understand the mechanisms and dynamics of the teaming process better.

We choose the duration of a match in seconds as the dependent variable. The independent variables are the metrics associated with whichever team won as well as the differences (Δ_W) in metrics between the winning team and the losing team. We use linear regression to estimate the effects of these metrics on the duration of the game. With respect to Table 5.2, we focus on the rows of the table whereas the duration represents the dependent variable. More particularly, we focus on the first row as we here always take the perspective of the winning team.

Descriptive Statistics

The correlations between all dependent and independent variables are listed in Table 5.8. The skill measures of a winning team are moderately correlated. Also the number of co-play relations is moderately correlated with the skill measures. The number of unique partners of a winning team is moderately correlated with the winning team kills and deny rate and weakly correlated with its death rate and the number of co-play relations. The differential measures are all moderately correlated with the corresponding main effect measure, i.e., winner team kills with Δ_W team kills, winner death rate with Δ_W death rate, winner deny rate with Δ_W deny rate, winner co-play with Δ_W co-play, and winner partners with Δ_W partners. The duration of a match is weakly correlated with most of the variables. Whether or not Radiant wins is weakly negatively correlated to the difference in the number of co-play relations of the winning team and the losing team;

i.e., if the winning team has more co-play relations than the losing team it is less likely that Radiant wins.

Models and Results

Table 5.9 shows the two Models. In Model 1 we include the compositional, relational and team ecosystem metrics of the winning team. We also add Radiant wins as a control variable, to capture which team succeeded. In Model 2, we include the differences in the metrics for the difference in compositional, relational and ecosystem factors between the winning and losing team.

The results for Model 1 indicate that overall team Radiant tends to spend 51.54 seconds less to win than team Dire; according to Model 2 it is 60.13 seconds. Teams with higher team kills spend more time to win and a higher team kills of one in both teams leads to an 8.99 seconds longer match when the difference between the teams remains constant (i.e., winner team kills is increased by one and Δ_W team kills remains the same). On the other hand, the differential advantage over the opposing team reduces the winning time and one kills advantage leads to a 5.86 seconds shorter match. Teams with a higher death rate spend more time to win. A higher death rate of one in both teams (i.e., winner death rate is increased by one and Δ_W death rate remains the same) in a match leads to a 28.38 seconds longer match; if death rate of a team increases by one compared to the other team (i.e., Δ_W death rate is increased by one and winner death rate remains the same), the team needs 14.60 seconds more to win. Teams with a higher deny rate spend less time to win. A team that has a higher deny rate of one compared to the other team spends 19.35 seconds less to win. In the second model, only the differential effect is significant, not the main effect. Each additional co-play relation on the winning team (i.e., winner co-play) leads to a 12.20 seconds longer match when the difference to the losing team (i.e., Δ_W co-play) stays the same; Each additional co-play relation the winning team has more than the losing team (i.e., Δ_W co-play) results in 13.86 seconds less to win. Winning teams with external unique partners spend less time to win but the differential advantage increases the time it takes to win.

5.2.5 Findings and Conclusions

In this analysis we explore the impacts of different types of team factors on performance and duration of matches in a team-vs-team setting. These factors are related to player's skills (compositional or individual level), co-play relations (relational or network level) and the familiarity with other players than the own teammates (team ecosystem or group level). We propose that teams are more likely to win if they have higher players' skills (H1), if they share more previous co-playing experiences (H2) and if they comprise players who played in many different teams before (H3). To test these hypotheses we introduce five match categories. These categories are based on the distribution of the duration of the matches. For short, medium low and medium duration matches all hypotheses are supported. However, relations tend to have a stronger impact for short duration matches whereas skills are especially important for medium duration matches. For longer matches

	Duration	Radiant Wins	W Team kills	W Death rate	W Deny rate	W Co-play	W Partners	Δ_W Team kills	Δ_W Death rate	Δ_W Deny rate	Δ_W Co-play
Duration	1.00										
Radiant Wins	-0.04***	1.00									
W Team kills	-0.03***	0.02***	1.00								
W Death rate	0.01**	0.01***	0.56***	1.00							
W Deny rate	-0.06***	0.02***	0.59***	0.55***	1.00						
W Co-play	0.00	-0.13***	0.43***	0.35***	0.38***	1.00					
W Partners	-0.14***	0.05***	0.46***	0.25***	0.50***	0.17***	1.00				
Δ_W Team kills	-0.01**	0.03***	0.46***	0.18***	0.09***	0.15***	0.05***	1.00			
Δ_W Death rate	0.01**	0.01***	0.15***	0.56***	0.23***	0.14***	0.06***	0.31***	1.00		
Δ_W Deny rate	-0.02***	0.00	0.10***	0.25***	0.52***	0.10***	0.05***	0.20***	0.43***	1.00	
Δ_W Co-play	-0.01*	-0.21***	0.13***	0.14***	0.09***	0.64***	-0.04***	0.25***	0.23***	0.15***	1.00
Δ_W Partners	-0.04***	0.06***	0.09***	0.09***	0.10***	-0.04**	0.48***	0.13***	0.13***	0.13***	-0.11***

Note: *p<0.05; **p<0.01; ***p<0.001; *"Winner" is abbreviated as "W".*

Table 5.8: Correlation of variables in duration models.

	Dependent variable:	
	Duration	
Measures	Model 1	Model 2
Radiant Wins	-51.54***	-60.13***
	(5.37)	(5.43)
Winner Team kills	5.60***	8.99***
	(1.48)	(1.94)
Winner Death rate	44.60***	28.38***
	(4.77)	(6.10)
Winner Deny rate	-19.35**	-0.91
	(5.92)	(8.01)
Winner Co-play	-0.42	12.20***
	(1.66)	(2.27)
Winner Partners	-5.97***	-7.51***
	(0.19)	(0.24)
Δ_W Team kills		-5.86***
		(1.58)
Δ_W Death rate		14.60**
		(4.64)
Δ_W Deny rate		-25.20***
		(6.22)
Δ_W Co-play		-13.86***
		(1.74)
Δ_W Partners		2.23***
		(0.24)
Intercept	2825.88***	2822.19***
	(6.65)	(7.40)
Pseudo R^2	0.023	0.027
N	62,034	62,034

Table 5.9: Models for duration.

the outcome is basically unpredictable. They might represent tie situations where random effects decide on winning and losing. Alternatively, they might be explained by additional variables not included in our model.

We find different patterns of impact mechanisms of the three performance factors on the time it takes for a team to win. The compositional and relational factors have the same patterns: high levels of skills and previous collaboration in the winning team without differential advantages lead to a longer match and differential advantages of one team over the other lead to a shorter match. On the other hand, the team ecosystem factor has an opposite pattern: when members of the winning team played in many different teams, they tend to finish the game faster; matches where the winning team has a higher differential of external partners will take a bit longer to win.

Both the relative performance as well as the duration models fix one of the two dimension that we are interested in, i.e., the duration is fixed in the first case and the results in

the latter. However, it is still an open issue, which we will address in future work, to model both duration and results at the same time as they constitute two measures of one outcome. Furthermore, we will enrich the models by including information about the heroes a team is composed of. In order to achieve this, we plan to apply data mining techniques. However, the heroes are very complex in Dota 2 and hard to classify.

5.3 Discussion

In this chapter we demonstrate some straightforward ways to integrate all levels of information into one model. Basically, we construct covariate variables related to the different levels and combine them either within a regression model or a network model. As discussed in Section 5.1, in the context of regression models one has to pay particularly attention when introducing covariate variables based on network structure as independence assumptions might get violated. In the context of the team-vs-team setting discussed in Section 5.2 this is less an issue as the teams are separated entities, at least from our simplifying perspective. The inter-teams relations that we consider refer to an earlier state of the system, i.e., before the observation period December 8^{th} to 14^{th}, 2011. Such a longitudinal setting facilitates the construction of covariates for regression models in a correct way [Sni11]. However, in future work we plan to develop models to address inter-team dependencies explicitly, e.g., with the help of Conditional Random Field Models (see Section 2.2.4). Such a model would comprise two distinct network levels.

From an application-oriented point of view, taking into account covariates from different levels enable to make use of richer information. Furthermore, one can compare the importance of covariates related to distinct levels. As discussed in the team-vs-team setting, for instance, relations within a team are as important as the skills of the players to explain the results. This helps to gain better insights into the dynamics of an empirical setting. In the team-vs-team setting the influence mechanisms are very hard to capture, in particular as we study the behavior of a group of people rather than individuals.

Furthermore, in our future work, we aim to explore further ways to integrate the different levels of information. One method that we consider is related to hyperbolic embedding [KPK$^+$10]. This approach is used in computer science to tackle certain problems, e.g., Internet routing. It has been shown that the embedded networks have certain properties that facilitate the integration of the group and the network level [KPK$^+$10]: All nodes of a network are mapped into a hidden metric space; the distances in this space are related to node similarities; and more similar/close nodes are more likely to be connected. This implies the following: More similar nodes are closer in the space; links between nodes exist with a probability that decreases with hidden distance; and more similar/close nodes are more likely to be connected [HN15].

CHAPTER 6

Conclusions

The aim of this thesis was to examine social influence mechanisms in order to integrate such processes into computational models on a large scale. Social influence is defined as a change of a person's behavior (or attitudes or beliefs, etc.) according to the behavior of other people in the social system.

As this topic is approached by a number of distinct disciplines including sociology, psychology, marketing and computer science that all have their own terminologies and methods, we started with an exploratory literature review. This resulted in our research contribution **M1**, i.e., an ontological framework to integrate and categorize different approaches to social influence mechanisms in a systematic way. This *Social Influence Framework* distinguishes three levels of information, i.e., the individual level, the group level and the network level. At the first level the individuals are considered as independent, at the second level groups of individuals with similar characteristics are considered. Here, it is not of interest whether or not the individuals interact. However, this is the focus of the network level, where interactions among individuals are emphasized. In general, social influence mechanisms can either be communication-based or comparison-based. The first refers to behavioral changes through direct interactions, the latter to behavioral changes due to observations of others who are perceived as relevant, e.g., who have similar characteristics or are in the same social position. The literature review resulted in the following research questions:

- **Research Question 1 (RQ1)**: Are there computational frameworks that integrate different levels and approaches when studying social influence on a large scale?
- **Research Question 2 (RQ2)**: What do we gain by taking different levels of information into account when studying social influence phenomena?

As it is an empirical question, whether or not social influence occurs, we distinguished two layers within our studies: The first layer comprises mathematical, statistical and

computational models to capture social influence mechanisms. The second layer is about the application of the models to empirical questions in order to obtain concrete statements about different domains. This was also how we organized our analysis.

First we studied social influence processes at the group level. A dataset from literature on teenage smoking behavior was used to illustrate the main concepts as well as the differences to the individual level. After that we applied the GDA approach to introduce a new perspective on user modeling in the context of personality-based recommender systems. We demonstrated how collective preferences can be used to describe users and their travel behavioral patterns as a result of both personality traits and social context. Seven basic dimensions obtained with the help of factor analysis represented these travel behavioral patterns and formed the basis of a metric space. The position of a user in this metric space captured her preferences and thus determined her user model. To elicit the preferences of a user, we introduced an innovative picture-based approach. Statistical analyses showed that these user models are both meaningful and capable of representing the setting in an accurate way. With the help of cluster analysis, groups were detected in which the users exhibited normative behavior. These groups can in turn be targeted by recommender systems. Thus, the results **M2**, i.e., in-depth statistical analyses of the setting, are contributions to the picture-based approach to recommender systems. Together with this various concrete insights into travel preferences of different user groups were obtained. These statements form our contribution **A1**. The social influence mechanisms that occur in this context are comparison processes.

Next we focused on the network level. We described the conventional social influence models LNAMs and ALAAMs for cross-sectional network data. Furthermore, we introduced CRF Models as a novel way to capture social influence in networks. This is our research contribution **M3**. One major advantage of CRF Models is that they can handle very well big networks and a high number of predictor variables, as opposed to the conventional models. Again we used the example of the smoking teenagers to illustrate the main ideas and to show the differences between the network level and the individual level. The conventional models as well as the introduced CRF Model were used in an empirical study about churn behavior in the MMORPG EverQuest II. Based on literature, hypotheses related to the commitment of a player, to her achievements and to social effects were phrased and tested that led to various insights into the motivations of players to quit. These insights represent our results **A2**.

Another model that we introduced was also related to the group level. Here the aim was combining structure and content of user discussion in online forums. We utilized LNAMs to study interdependencies between sentiments of the users of the forum. To obtain these sentiments we applied text mining techniques and sentiment analysis to free-text comments of the users. Thus, these models, our contribution **M4**, represent a novel way to study sentiment contagion in networks. We apply this model in the context of an online travel forum. The results imply that the emotions are in fact interrelated. Furthermore we find out that also individual attributes such as gender have an impact

on the sentiments of a user. These are our results **A3**. On the network level, we focused on communication based social influence mechanism.

Finally, we examined approaches how all the three levels could be integrated into one model. In the case of the teenage smoking behavior we used information of the group level and included it as nodal attribute in the CRF model. Next we studied a considerably complex setting related to team-vs-team competitions. Here the winning behavior or performance of a team comprised two dimensions, the result of the match in terms of winning or losing, on the one hand, and the duration of a match, on the other hand. Furthermore, the performance of a team had to be considered relative to the performance of its opponent. To address the dynamics of this complex setting, factors related to distinct levels of information were constructed. We introduced two types of models, i.e., relative outcome models and duration models, that integrated those factors and that were indirectly able to account for both outcome dimensions. These models are our research contribution **M5**. With the help of these models we analyzed relative performance of teams in the MOBA game Dota 2. The statements about the relative importance of the skills of the members of a team, i.e., factors related to the individual level, of past co-play experience among the team members, i.e., a factor related to the network level, and of having a lot of experiences of playing in other teams, i.e., a factor related to the group level, provide concrete insights into preconditions that might help a team to win quickly. These statements are our results **A4**. Here, communication-based social influence occurs, i.e., the co-play relations, as well as comparison-based social influence. Players who were parts of many different teams before know a lot of different play-styles and have seen various tricks. However, the behavior that we studied here, is the behavior of a group of people, which is obviously even more complex.

Now, getting back to the research questions **RQ1** and **RQ2**. We can relate our results **M1**, **M2**, **M3**, **M4** and **M5** to **RQ1** as they all represent successful attempts to study the impact of social context on human behavior by integrating different levels of information (whereas **M1** provides the basis for doing so). In all the presented models individual attributes representing the individual level were taken into account, i.e., either when analyzing the metric space or when studying social networks. In the context of the team-vs-team setting we were able to develop models that combined all three levels of information. Furthermore, we captured a two dimensional outcome behavior.

Except for the LNAM-based approach, i.e., the network models combining structure and contents of discussion (see **M4**), all presented models are capable of handling large-scale settings. In particular CRF Models are very well suited for these kind of applications. Just as a comparison, as we saw in Chapter 4, it took two days to fit an ALAAM to the churn network, two hours to fit a LNAM but less than one minute for the CRF model.

The answer to **RQ2** is provided by the results **A1**, **A2**, **A3** and **A4**. In each of these studies we are able to obtain a more comprehensive picture by taking different levels into account. This we showed within the different studies, in particular by comparing their outcomes to simple regression models.

In this thesis we presented various perspectives on social influence processes and saw that this is a very complex and multi-level topic facing a lot of challenges. In our work we started to approach the problem of modeling social influence phenomena on a large-scale while integrating different levels and aspects. We were able to meet some of the challenges. However, there are still various open issues to address.

As social influence is an inherently interdisciplinary topic, the literature review that we presented is by no means exhaustive. Thus, we will continue to look for relevant literature in order to extend the *Social Influence Framework*.

Regarding **RQ1**, we are particularly interested in finding and improving ways to integrate the different levels of information. Ideally, these models should be applicable to large-scale settings and preserve the advantages of each level.

We proposed CRF Models as a novel way to study an outcome behavior while integrating the individual and the network level. In Chapter 4 we showed in different application scenarios that CRF Models provide a good alternative to both logistic regression as well as statistical network models for cross-sectional data, i.e., LNAMs and ALAAMS. In particular, the ALAAMs are very complex as there the joint distribution of network and behavior is modeled, whereas we consider the network as fixed and study the behavior conditioned on the network. Thus, ALAAMs are more general and computational very expensive. What is needed next is a more systematic way of comparing the different models, i.e., by using simulated data. For such data we can specify the strength of the influence mechanisms; thus, we know the ground truth. This enables to assess which model is the most accurate. It can be studied, moreover, whether in fact social influence is detected or rather social selection. Furthermore, we will also compare the models theoretically to better access how exactly ALAAMs and CRF Models are related. Another open issue that we are working on is to find a way to compute the standard error of the edge parameters of the CRF Models. Here, a bootstrap method might be an option. Once the evidence is provided that CRF Models are an effective and efficient way to capture and identify social influence mechanism, we plan to provide this method as a software package that can be used for statistical inference in social influence models.

In this work we did not explicitly model comparison mechanisms at the network level, i.e., part D in Figure 1.1. However, this is also planned for future work. There is some work that uses the concept of structural equivalence to study comparison mechanism in this context [Lee02]. As we aim to study online social networks also the concepts of automorphic equivalence and regular equivalence might be appropriate (see Section 2.2.3 for their definitions). However, it might not be straightforward to operationalize these concepts.

We will also continue to develop and improve approaches to integrate all three levels of information. In particular we are strongly interested in supplementing our recommender model by structural information in order to take social influence of direct interaction of a user, i.e., communication-based influence, into account when delivering recommendations. As discussed at the end of Chapter 5, hyperbolic embedding might be one option to

combine the GDA approach with statistical network models and thus integrate group and network level [KPK+10].

The team-vs-team models introduced in Chapter 5 simplify the complex head-to-head setting in different regards: First, the models account for two outcome dimensions by fixing either the results or the durations of the matches. However, our aim is to find models that account for both dimensions at the same time. Second, the ecosystem measure reduces inter-team relations to a simple number. It would be more correct to model the setting in a more complex way, e.g., with the help of hypergraphs, i.e., a graph in which an edge can connect more than two nodes.

Advances with respect to the outlined plans for future work will automatically lead to additional insights regarding **RQ2**.

List of Figures

List of Tables

Bibliography

[ABC09] Deborah Ancona, Henrik Bresman, and David Caldwell. The x-factor:: Six steps to leading high-performing x-teams. *Organizational dynamics*, 38(3):217–224, 2009.

[Abd03] Hervé Abdi. Factor rotations in factor analyses. *Encyclopedia for Research Methods for the Social Sciences. Sage: Thousand Oaks, CA*, pages 792–795, 2003.

[AC92] Deborah G. Ancona and David F. Caldwell. Bridging the boundary: External activity and performance in organizational teams. *Administrative Science Quarterly*, 37(4):634–665, 1992.

[Agg11] Charu C. Aggarwal. *An Introduction to Social Network Data Analytics*, book section 1, pages 1–15. Springer US, 2011.

[Ans88] Luc Anselin. *Spatial econometrics: methods and models*, volume 4. Springer Science and Business Media, 1988.

[AU07] Luis A Nunes Amaral and Brian Uzzi. Complex systems-a new paradigm for the integrative study of management, physical, and technological systems. *Management Science*, 53(7):1033–1035, 2007.

[Bak07] Gökhan Bakir. *Predicting structured data*. MIT press, 2007.

[BBJ97] T. T. Baldwin, M. D. Bedell, and J. L. Johnson. The social fabric of a team-based m.b.a. program: Network effects on student satisfac-tion and performance. *Academy of Management Journal*, 40(6):1369–1397, 1997.

[BCM11] Smriti Bhagat, Graham Cormode, and S. Muthukrishnan. *Node classification in social networks*, pages 115–148. Springer, 2011.

[BD82] Ronald S. Burt and Patrick Doreian. Testing a structural model of perception: Conformity and deviance with respect to journal norms in elite sociological methodology. *Quality and Quantity*, 16(2):109–150, 1982.

[BDD+07] Helmut Berger, Michaela Denk, Michael Dittenbach, Dieter Merkl, and Andreas Pesenhofer. Quo vadis homo turisticus? towards a picture-based tourist profiler. *Information and Communication Technologies in Tourism 2007*, pages 87–96, 2007.

[BES10] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[BH06] Prasad Balkundi and David A. Harrison. Ties, leaders, and time in teams: Strong inference about network structure's effects on team viability and performance. *Academy of Management Journal*, 49(1):49–68, 2006.

[BKL09] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python.* " O'Reilly Media, Inc.", 2009.

[BM14] Ronald S Burt and Jennifer Merluzzi. Embedded brokerage: Hubs versus locals. *Research in the Sociology of Organizations: Contemporary Perspectives on Organizational Social Networks, edited by DJ Brass, G. Labianca, A. Mehra, DS Halgin, and SP Borgatti*, 40:161–177, 2014.

[Bor64] Edgar F. Borgatta. A very short test of personality: the behavioral self-rating (bsr) form. *Psychological Reports*, 14(1):275–284, 1964.

[Bou84] Pierre Bourdieu. *Distinction: A social critique of the judgement of taste.* Harvard University Press, 1984.

[BP12] Sergey Brin and Lawrence Page. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, 56(18):3825–3833, 2012.

[BR11] Robin D. Burke and Maryam Ramezani. Matching recommendation technologies and domains. *Recommender Systems Handbook*, 1:367, 2011.

[BS13] Jörg Blasius and Andreas Schmitz. *Sozialraum-und Habituskonstruktion Die Korrespondenzanalyse in Pierre Bourdieus Forschungsprogramm*, pages 201–218. Springer, 2013.

[Bur87] Ronald S. Burt. Social contagion and innovation: Cohesion versus structural equivalence. *American journal of Sociology*, pages 1287–1335, 1987.

[Bur00] Ronald S. Burt. The network structure of social capital. *Research in organizational behavior*, 22:345–423, 2000.

[Bü10] Achim Bühl. *SPSS 18 (ehemals PASW): Einführung in die moderne Datenanalyse*, volume 4028. Tata McGraw-Hill Education, 2010.

[Car10] Carter T. Butts. sna: Tools for Social Network Analysis. `https://cran.r-project.org/web/packages/sna`, 2010.

[CB97] Susan G. Cohen and Diane E. Bailey. What makes teams work: Group effectiveness research from the shop floor to the executive suite. *Journal of management*, 23(3):239–290, 1997.

[CCH+08] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 160–168. ACM, 2008.

[CdGF+13] Li Chen, Marco de Gemmis, Alexander Felfernig, Pasquale Lops, Francesco Ricci, and Giovanni Semeraro. Human decision making and recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(3):17, 2013.

[CG04] Robert B. Cialdini and Noah J. Goldstein. Social influence: Compliance and conformity. *Annual Review of Psychology*, 55(1):591–621, 2004.

[CG13] Marcus Carter and Martin R. Gibbs. esports in eve online: Skullduggery, fair play and acceptability in an unbounded competition. In *FDG*, pages 47–54, 2013.

[CH02] Pei-Yu Chen and Lorin M Hitt. Measuring switching costs and the determinants of customer retention in internet-enabled businesses: A study of the online brokerage industry. *Information Systems Research*, 13(3):255–274, 2002.

[CH12] Gifford Cheung and Jeff Huang. Remix and play: lessons from rule variants in texas hold'em and halo 2. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 559–568. ACM, 2012.

[Cho15] Antonios Chorianopoulos. *Effective CRM using Predictive Analytics*. Wiley, 2015.

[CK87] Robert A. Cooke and John A. Kernaghan. Estimating the difference between group versus individual performance on problem-solving tasks. *Group and Organization Management*, 12(3):319–342, 1987.

[Coh74] Erik Cohen. Who is a tourist?: A conceptual clarification1. *The sociological review*, 22(4):527–555, 1974.

[Con85] James J. Conley. Longitudinal stability of personality traits: A multitrait–multimethod–multioccasion analysis. *Journal of personality and social psychology*, 49(5):1266, 1985.

[Cor14] Valve Corporation. Valve software, 2014.

[CPD15] Rafael A. Calvo, Dorian Peters, and Sidney D'Mello. When technologies manipulate our emotions. *Communications of the ACM*, 58(11):41–42, 2015.

[CS14] John Cannarella and Joshua A. Spechler. Epidemiological modeling of online social network dynamics. *arXiv preprint arXiv:1401.4208*, 2014.

[CSS+11] Anna Chmiel, Pawel Sobkowicz, Julian Sienkiewicz, Georgios Paltoglou, Kevan Buckley, Mike Thelwall, and Janusz A. Hołyst. Negative emotions boost user activity at bbc forum. *Physica A: statistical mechanics and its applications*, 390(16):2936–2944, 2011.

[CST+11] Anna Chmiel, Julian Sienkiewicz, Mike Thelwall, Georgios Paltoglou, Kevan Buckley, Arvid Kappas, and Janusz A. Hołyst. Collective emotions online and their influence on community life. *PloS one*, 6(7):e22207, 2011.

[CT98] Robert B. Cialdini and Melanie R. Trost. *Social influence: Social norms, conformity and compliance*, volume 2, page 151–192. Boston: McGraw-Hill, 4th edition, 1998.

[CVdP08] Kristof Coussement and Dirk Van den Poel. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert systems with applications*, 34(1):313–327, 2008.

[DMML00] Piew Datta, Brij Masand, DR Mani, and Bin Li. Automated cellular modeling and prediction on a large scale. *Artificial Intelligence Review*, 14(6):485–502, 2000.

[dMSS+14] Yves-Alexandre de Montjoye, Arkadiusz Stopczynski, Erez Shmueli, Alex Pentland, and Sune Lehmann. The strength of the strongest ties in collaborative problem solving. *Scientific reports*, 4, 2014.

[DNW16] Amra Delic, Julia Neidhardt, and Hannes Werthner. Are sun lovers nervous? In *ENTER 2016*, volume 7, page 5. e-Review of Tourism Research (eRTR), 2016.

[Dor89] Patrick Doreian. Network autocorrelation models: Problems and prospects. *Spatial Statistics: Past, Present, Future. Ann Arbor, Michigan Document Services*, 1989.

[DR13] Galina Daraganova and Garry Robins. Autologistic actor attribute models. *Exponential Random Graph Models for Social Networks: Theoiy, Methods and Applications, Lusher, D., J. Koskinen and G. Robins (Eds.). CambridgeUniversity Press, New York*, pages 102–114, 2013.

[DSV+08] Koustuv Dasgupta, Rahul Singh, Balaji Viswanathan, Dipanjan Chakraborty, Sougata Mukherjea, Amit A Nanavati, and Anupam Joshi. Social ties and their relevance to churn in mobile telecom networks. In *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*, pages 668–677. ACM, 2008.

[dVSV14] Mathijs de Vaan, David Stark, and Balizs Vedres. Game changer. the topology of creativity. *Stato e mercato*, 34(3):307–340, 2014.

[DWA10] Jordi Duch, Joshua S. Waitzman, and Luís A Nunes Amaral. Quantifying the performance of individual players in a team activity. *PloS one*, 5(6):e10937, 2010.

[DYNM06] Nicolas Ducheneaut, Nicholas Yee, Eric Nickell, and Robert J Moore. Alone together?: exploring the social dynamics of massively multiplayer online games. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 407–416. ACM, 2006.

[EK10] David Easley and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[ES06] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer, 2006.

[Exp14] Expedia, Inc. `http://www.expedia.at`, April 2014.

[FBR99] Stacie Furst, Richard Blackburn, and Benson Rosen. Virtual team effectiveness: a proposed research agenda. *Information Systems Journal*, 9(4):249–269, 1999.

[FBS07] Wu-chang Feng, David Brandt, and Debanjan Saha. A long-term study of a popular mmorpg. In *Proceedings of the 6th ACM SIGCOMM Workshop on Network and System Support for Games*, pages 19–24. ACM, 2007.

[FCS10] David A. Freedman, David Collier, and Jasjeet S. Sekhon. *Statistical models and causal inference: a dialogue with the social sciences*. Cambridge University Press, 2010.

[FH12] Ian Fellows and Mark S. Handcock. Exponential-family random network models. *arXiv preprint arXiv:1208.0121*, 2012.

[FJ97] Noah E. Friedkin and Eugene C. Johnsen. Social positions in influence networks. *Social Networks*, 19(3):209–222, 1997.

[FNP+12] Anna Fensel, Julia Neidhardt, Nataliia Pobiedina, Dieter Fensel, and Hannes Werthner. Towards an intelligent framework to understand and feed the web. In *Business Information Systems Workshops*, pages 255–266. Springer, 2012.

[FRC59] John R.P. French, Bertram Raven, and D. Cartwright. The bases of social power. *Classics of organization theory*, pages 311–320, 1959.

[Fri98] Noah E. Friedkin. *A structural theory of social influence*, volume 13. Cambridge University Press, 1998.

[Fri01] Noah E. Friedkin. Norm formation in social influence networks. *Social Networks*, 23(3):167–189, 2001.

[FST15] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. Using instagram picture features to predict users' personality. In *MultiMedia Modeling*, pages 850–861. Springer, 2015.

[FZCX14] Rui Fan, Jichang Zhao, Yan Chen, and Ke Xu. Anger is more influential than joy: Sentiment correlation in weibo. *PloS one*, 9(10):e110184, 2014.

[GD96] Richard A. Guzzo and Marcus W. Dickson. Teams in organizations: Recent research on performance and effectiveness. *Annual Review of Psychology*, 47(1):307–338, 1996.

[GF06] Ulrike Gretzel and Daniel R. Fesenmaier. Persuasion in recommender systems. *International Journal of Electronic Commerce*, 11(2):81–100, 2006.

[GH13] Shirley Gregor and Alan R. Hevner. Positioning and presenting design science research for maximum impact. *MIS quarterly*, 37(2):337–355, 2013.

[GJWL13] Lynn Gao, James Judd, Dave Wong, and Jamie Lowder. Classifying dota 2 hero characters based on play style and performance. Technical Report, November 2013.

[GMHF04] Ulrike Gretzel, Nicole Mitsche, Yeong-Hyeon Hwang, and Daniel R. Fesenmaier. Tell me who you are and i will tell you where to go: use of travel personalities in destination recommendation systems. *Information Technology and Tourism*, 7(1):3–12, 2004.

[Gol93] Lewis R. Goldberg. The structure of phenotypic personality traits. *American psychologist*, 48(1):26, 1993.

[GSO11] Inma Garcia, Laura Sebastia, and Eva Onaindia. On the design of individual and group recommender systems for tourism. *Expert systems with applications*, 38(6):7683–7692, 2011.

[GUSA05] Roger Guimera, Brian Uzzi, Jarrett Spiro, and Luis A. Nunes Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702, 2005.

[GY02] Heather Gibson and Andrew Yiannakis. Tourist roles: Needs and the lifecourse. *Annals of tourism research*, 29(2):358–383, 2002.

[GZFF12] Dietmar Gräbner, Markus Zanker, Gunther Fliedl, and Matthias Fuchs. *Classification of customer reviews based on sentiment analysis*. na, 2012.

[Hay02] Colin Hay. *Political analysis: a critical introduction*. Palgrave Macmillan, 2002.

[HFG+16] Dirk Helbing, Bruno S. Frey, Gerd Gigerenzer, Ernst Hafen, Michael Hagner, Yvonne Hofstetter, Jeroen van den Hoven, Roberto V. Zicari, and Andrej Zwitter. Digitale demokratie statt datendiktatur. *Spektrum der Wissenschaft*, 1, 2016.

[HMPR04] Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS quarterly*, 28(1):75–105, 2004.

[HN15] Yun Huang and Julia Neidhardt. From networks to space: Constructing metric spaces for social interactions. `http://www.ec.tuwien.ac.at/neidhardt/EISSII.pdf`, October 2015. Empirical Investigation of Social Space II, Bonn, Germany.

[HP10] Rong Hu and Pearl Pu. *A study on user perception of personality-based recommender systems*, pages 291–302. Springer, 2010.

[HR05] Robert A. Hanneman and Mark Riddle. Introduction to social network methods, 2005.

[HSH+08] James Hendler, Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, and Daniel Weitzner. Web science: an interdisciplinary approach to understanding the web. *Communications of the ACM*, 51(7):60–69, 2008.

[HTK97] Thorsten Hennig Thurau and Alexander Klee. The impact of customer satisfaction and relationship quality on customer retention: A critical reassessment and model development. *Psychology and Marketing*, 14(8):737–764, 1997.

[HTRR06] John Hadden, Ashutosh Tiwari, Rajkumar Roy, and Dymtr Ruta. Churn prediction using complaints data. In *Proceedings Of World Academy Of Science, Engineering and Technology*. Citeseer, 2006.

[HW10] Daniel M. Hausman and Brynn Welch. Debate: To nudge or not to nudge*. *Journal of Political Philosophy*, 18(1):123–136, 2010.

[HYW06] Shin-Yuan Hung, David C Yen, and Hsiu-Yu Wang. Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3):515–524, 2006.

[Inc14] Curse Inc. Dota 2 wiki, 2014.

[JR09] Aparna Joshi and Hyuntak Roh. The role of context in work team diversity research: A meta-analytic review. *Academy of Management Journal*, 52(3):599–627, 2009.

[JS99] Oliver P. John and Sanjay Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138, 1999.

[JZFF10] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.

[KD15] Clemens Költringer and Astrid Dickinger. Analyzing destination branding and image from online sources: A web content mining approach. *Journal of Business Research*, 2015.

[Kel58] Herbert C. Kelman. Compliance, identification, and internalization: Three processes of attitude change. *Journal of conflict resolution*, pages 51–60, 1958.

[KF09] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[KG99] Tally Katz-Gerro. Cultural consumption and social stratification: leisure activities, musical tastes, and social location. *Sociological Perspectives*, 42(4):627–646, 1999.

[KGH14] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[KKT03] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[Kle92] Paul Klemperer. Equilibrium product lines: Competing head-to-head may be less competitive. *The American Economic Review*, pages 740–755, 1992.

[KPK+10] Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguná. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.

[KPS09] Jaya Kawale, Aditya Pal, and Jaideep Srivastava. Churn prediction in mmorpgs: A social influence based approach. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 423–428. IEEE, 2009.

[KPX+15] Worarat Krathu, Christian Pichler, Guohui Xiao, Hannes Werthner, Julia Neidhardt, Marco Zapletal, and Christian Huemer. Inter-organizational success factors: a cause and effect model. *Information Systems and e-Business Management*, 13(3):553–593, 2015.

[KRC+11] Marcel Karnstedt, Matthew Rowe, Jeffrey Chan, Harith Alani, and Conor Hayes. The effect of user features on churn in social networks. In *Proceedings of the 3rd International Web Science Conference*, page 23. ACM, 2011.

[Kri12] Wolfgang Krinninger. Inspiration and information – critical factors to the success of travel platforms. Master's thesis, Vienna University of Economics and Business, November 2012.

[Lat96] Bibb Latané. Dynamic social impact: The creation of culture by communication. *Journal of Communication*, 46:13–25, 1996.

[Lee02] Roger T. A.J. Leenders. Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks*, 24(1):21–47, 2002.

[Liq14] LiquidDota - Dota 2 Forums. Radiant vs Dire By Duration. `http://www.liquiddota.com/forum/dota-2-general/460223-radiant-vs-dire-by-duration`, November 2014.

[LKR12] Dean Lusher, Johan Koskinen, and Garry Robins. *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press, 2012.

[LMP01] John D. Lafferty, Andrew McCallum, and Fernando C.N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, 2001.

[LPA+09] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, and Myron Gutmann. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[LRR04] Brigitte Le Roux and Henry Rouanet. *Geometric data analysis: from correspondence analysis to structured data analysis*. Springer Science and Business Media, 2004.

[Mar14] Stephen Marsland. *Machine learning: an algorithmic perspective*. CRC press, 2014.

[MC00] Martha L. Maznevski and Katherine M. Chudoba. Bridging space over time: Global virtual team dynamics and effectiveness. *Organization science*, 11(5):473–492, 2000.

[MC03] Peter R. Monge and Noshir S. Contractor. *Theories of communication networks*. Oxford University Press, 2003.

[MC14] Sears Merritt and Aaron Clauset. Scoring dynamics across professional team sports: tempo, balance and predictability. *EPJ Data Science*, 3(1):4, 2014.

[Mic97] Lynn Michell. Loud, sad or bad: young people's perceptions of peer groups and smoking. *Health Education Research*, 12(1):1–14, 1997.

[MPT10] Marija Mitrović, Georgios Paltoglou, and Bosiljka Tadić. Networks and emotion-driven user communities at popular blogs. *The European Physical Journal B*, 77(4):597–609, 2010.

[MR11] Lorraine McGinty and James Reilly. *On the evolution of critiquing recommenders*, pages 419–453. Springer, 2011.

[MSW+11] Mahalia Miller, Conal Sathi, Daniel Wiesenthal, Jure Leskovec, and Christopher Potts. Sentiment flow through hyperlink networks. In *ICWSM*, 2011.

[MW96] Lynn Michell and Patrick West. Peer pressure to smoke: the meaning depends on the method. *Health education research*, 11(1):39–49, 1996.

[New10] Mark Newman. *Networks: an introduction*. Oxford University Press, 2010.

[NHC15] Julia Neidhardt, Yun Huang, and Noshir Contractor. Team vs. team: Success factors in a multiplayer online battle arena game. In *Academy of Management Proceedings*, volume 2015, 2015.

[NHWC15] Julia Neidhardt, Yun Huang, Hannes Werthner, and Noshir Contractor. Conditional random field models as a way to capture peer influence in social networks. `http://www.ec.tuwien.ac.at/neidhardt/Sunbelt2015.pdf`, June 2015. Sunbelt XXXV, Brighton, UK.

[NPW15] Julia Neidhardt, Nataliia Pobiedina, and Hannes Werthner. What can we learn from review data? In *ENTER 2015*, volume 6, page 5. e-Review of Tourism Research (eRTR), 2015.

[NRW16] Julia Neidhardt, Nataliia Rümmele, and Hannes Werthner. Can we predict your sentiments by listening to your peers? In *Information and Communication Technologies in Tourism 2016*, pages 593–603. Springer International Publishing, 2016.

[NSSW14a] Julia Neidhardt, Rainer Schuster, Leonhard Seyfang, and Hannes Werthner. Eliciting the users' unknown preferences. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 309–312, 2645767, 2014. ACM.

[NSSW14b] Julia Neidhardt, Leonhard Seyfang, Rainer Schuster, and Hannes Werthner. A picture-based approach to recommender systems. *Information Technology and Tourism*, 15(1):49–69, 2014.

[OCL04] Hongseok Oh, Myung-Ho Chung, and Giuseppe Labianca. Group social capital and group effectiveness: The role of informal socializing ties. *Academy of management journal*, 47(6):860–875, 2004.

[PBNW11] Roland Piazzi, Rodolfo Baggio, Julia Neidhardt, and Hannes Werthner. Destinations and the web: a network analysis view. *Information Technology and Tourism*, 13(3):215–228, 2011.

[PBNW12] Roland Piazzi, Rodolfo Baggio, Julia Neidhardt, and Hannes Werthner. Network analysis of the austrian etourism web. In Matthias Fuchs, Francesco Ricci, and Lorenzo Cantoni, editors, *ENTER 2012, Information and Communication Technologies in Tourism*, pages 356–367. Springer Vienna, 2012.

[PC86] Richard E. Petty and John T. Cacioppo. *The elaboration likelihood model of persuasion*. Springer, 1986.

[Pea82] Philip L Pearce. *The social psychology of tourist behaviour: International Series in Experimental Social Psychology*, volume 3. Pergamon Press, 1982.

[Pix14] Pixtri OG. `http://www.pixmeaway.com`, April 2014.

[PNCM+13] Nataliia Pobiedina, Julia Neidhardt, Maria del Carmen Calatrava Moreno, Laszlo Grad-Gyenge, and Hannes Werthner. On successful team formation: Statistical analysis of a multiplayer online game. In *Business Informatics (CBI), 2013 IEEE 15th Conference on*, pages 55–62. IEEE, 2013.

[PNCMW13] Nataliia Pobiedina, Julia Neidhardt, Maria del Carmen Calatrava Moreno, and Hannes Werthner. Ranking factors of team success. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1185–1194. International World Wide Web Conferences Steering Committee, 2013.

[PUM+13] Chitra Phadke, Huseyin Uzunalioglu, Veena B Mendiratta, Dan Kushnir, and Derek Doran. Prediction of subscriber churn using social network analysis. *Bell Labs Technical Journal*, 17(4):63–75, 2013.

[PW03] Michael Pearson and Patrick West. Drifting smoke rings. *Connections*, 25(2):59–76, 2003.

[R F16] R Foundation. The R project for statistical computing. `http://www.r-project.org/`, January 2016.

[Ras07] Lisa Rashotte. *Social influence*, page 4426–4429. Blackwell Publishing, 2007.

[REP01] Garry Robins, Peter Elliott, and Philippa Pattison. Network models for social selection processes. *Social networks*, 23(1):1–30, 2001.

[RN07] Francesco Ricci and Quang Nhat Nguyen. Acquiring and revising preferences in a critique-based mobile recommender system. *Intelligent Systems, IEEE*, 22(3):22–29, 2007.

[RPE01] Garry Robins, Philippa Pattison, and Peter Elliott. Network models for social influence processes. *Psychometrika*, 66(2):161–189, 2001.

[RRS11] Francesco Ricci, Lior Rokach, and Bracha Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[RSW+07] Garry Robins, Tom Snijders, Peng Wang, Mark Handcock, and Philippa Pattison. Recent developments in exponential random graph (p*) models for social networks. *Social networks*, 29(2):192–215, 2007.

[RWE+09] Shoba Ramanadhan, Jean L. Wiecha, Karen M. Emmons, Steven L. Gortmaker, and Kasisomayajula Viswanath. Extra-team connections for knowledge transfer between staff teams. *Health Education Research*, 24(6):967–976, 2009.

[RWZ05] Francesco Ricci, Karl Woeber, and Andreas Zins. Recommendations by collaborative browsing. *Information and Communication Technologies in Tourism 2005*, pages 172–182, 2005.

[Sch16] Mark Schmidt. Ugm: Matlab code for undirected graphical models. `https://www.cs.ubc.ca/ schmidtm/Software/UGM.html`, February 2016.

[SDN15] Dimitris Sacharidis, Amra Delic, and Julia Neidhardt. Learning the role and behavior of users in group decision making. In *Proceedings of the 2nd International Workshop on Decision Making and Recommender Systems*, volume 1533, pages 25–28, 2015.

[Sha16] Cosma Rohilla Shalizi. Advanced data analysis from an elementary point of view. To be published by Cambridge University Press; Pre-print available at `http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV`, January 2016.

[SHFL13] Sergej Schmunk, Wolfram Höpken, Matthias Fuchs, and Maria Lexhagen. *Sentiment analysis: Extracting decision-relevant knowledge from UGC*, pages 253–265. Springer, 2013.

[SM08] Andreas Schwab and Anne S Miner. Learning in hybrid-project systems: The effects of project performance on repeated collaboration. *Academy of Management Journal*, 51(6):1117–1149, 2008.

[SM11] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Machine Learning*, 4(4):267–373, 2011.

[Sni11] Tom A. B. Snijders. Statistical models for social networks. *Annual Review of Sociology*, 37(1):131–153, 2011.

[Son16] Sony Online Entertainment. Everquest II. `https://www.everquest2.com/home`, January 2016.

[SPRH06] Tom A.B. Snijders, Philippa E. Pattison, Garry L. Robins, and Mark S. Handcock. New specifications for exponential random graph models. *Sociological methodology*, 36(1):99–153, 2006.

[SSP10] Christian Steglich, Tom A.B. Snijders, and Michael Pearson. Dynamic networks and behavior: Separating selection from influence. *Sociological methodology*, 40(1):329–393, 2010.

[SSW06] Christian Steglich, Tom A.B. Snijders, and Patrick West. Applying siena: An illustrative analysis of the co-evolution of adolescents' friendship networks, taste in music, and alcohol consumption. *Methodology*, 2(1):48–56, 2006.

[ST09] Cass R. Sunstein and Richard H. Thaler. *Nudge: Improving Decisions About Health, Wealth and Happiness*. Penguin UK, 2009.

[ST11] Jimeng Sun and Jie Tang. *A survey of models and algorithms for social influence analysis*, pages 177–214. Springer, 2011.

[Sta16] Statista. Facebook's advertising revenue worldwide from 2009 to 2015 (in billion U.S. dollars). `http://www.statista.com/statistics/271258/facebooks-advertising-revenue-worldwide/`, February 2016.

[TBT+11] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[TC11] Konstantinos K. Tsiptsis and Antonios Chorianopoulos. *Data Mining Techniques in CRM: Inside Customer Segmentation*. Wiley, 2011.

[THC+15] Thanassis Tiropanis, Wendy Hall, Jon Crowcroft, Noshir Contractor, and Leandros Tassiulas. Network science, web science, and internet science. *Communications of the ACM*, 58(8):76–82, 2015.

[Tom15a] Tom A.B. Snijders. Description 'Teenage Friends and Lifestyle Study' data. https://www.stats.ox.ac.uk/~snijders/siena/Glasgow_data.htm, January 2015.

[Tom15b] Tom A.B. Snijders. The Siena webpage. http://www.stats.ox.ac.uk/~snijders/siena/, January 2015.

[Tri14] Trip Technologies Inc. http://www.tripbase.com, April 2014.

[TWU10] Mike Thelwall, David Wilkinson, and Sukhvinder Uppal. Data mining emotion in social network communication: Gender differences in myspace. *Journal of the American Society for Information Science and Technology*, 61(1):190–199, 2010.

[Val16] Valve Corporation. Dota 2 - official blog. http://www.dota2.com/, January 2016.

[VdPL04] Dirk Van den Poel and Bart Lariviere. Customer attrition analysis for financial services using proportional hazard models. *European journal of operational research*, 157(1):196–217, 2004.

[VS13] Mike Vuolo and Jeremy Staff. Parent and child cigarette use: A longitudinal, multigenerational study. *Pediatrics*, 132(3):e568–e577, 2013.

[Was13] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science and Business Media, 2013.

[WASC+15] Hannes Werthner, Aurkene Alzua-Sorzabal, Lorenzo Cantoni, Astrid Dickinger, Ulrike Gretzel, Dietmar Jannach, Julia Neidhardt, Birgit Pröll, Francesco Ricci, Miriam Scaglione, Brigitte Stangl, Oliviero Stock, and Markus Zanker. Future research issues in it and tourism. *Information Technology and Tourism*, 15(1):1–15, 2015.

[WCP+11] Dmitri Williams, Noshir Contractor, Marshall Scott Poole, Jaideep Srivastava, and Dora Cai. The virtual worlds exploratorium: using large-scale data and computational techniques for communication research. *Communication Methods and Measures*, 5(2):163–180, 2011.

[WDX+06] Dmitri Williams, Nicolas Ducheneaut, Li Xiong, Yuanyuan Zhang, Nick Yee, and Eric Nickell. From tree house to barracks the social life of guilds in world of warcraft. *Games and culture*, 1(4):338–361, 2006.

[WF94] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[WHG01] Katharina M. Wurst, Martin M. Högl, and Hans G. Gemünden. Collaboration within and between teams in multi-team r&d projects. In *Management of Engineering and Technology, 2001. PICMET '01. Portland International Conference on*, volume Supplement, pages 545–552 vol.2, 2001.

[Wil10] Dmitri Williams. The mapping principle, and a research framework for virtual worlds. *Communication Theory*, 20(4):451–470, 2010.

[WK99] Hannes Werthner and Stefan Klein. *Information technology and tourism: a challenging ralationship*. Springer-Verlag Wien, 1999.

[Woo00] Wendy Wood. Attitude change: Persuasion and social influence. *Annual review of psychology*, 51(1):539–570, 2000.

[WRP06] Peng Wang, Garry Robins, and Philippa Pattison. Pnet: Program for the estimation and simulation of p* exponential random graph models, user manual. *Department of Psychology, University of Melbourne*, 2006.

[WRS02] Amy B. Woszczynski, Philip L. Roth, and Albert H. Segars. Exploring the theoretical foundations of playfulness in computer interactions. *Computers in Human Behavior*, 18(4):369–388, 2002.

[WS13] Thomas Weiss and Sabrina Schiele. Virtual worlds in competitive contexts: Analyzing esports consumer needs. *Electronic Markets*, 23(4):307–316, 2013.

[WSZ+06] Hannes Werthner, Oliviero Stock, Massimo Zancanaro, Daniel R. Fesenmaier, and Karl W. Wöber. Futuring travel destination recommendation systems. *Destination Recommendation Systems: Behavioral Foundations and Applications*, pages 297–314, 2006.

[YG92] Andrew Yiannakis and Heather Gibson. Roles tourists play. *Annals of tourism Research*, 19(2):287–303, 1992.

[ZAA07] Jun Zhang, Mark S. Ackerman, and Lada Adamic. Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, pages 221–230. ACM, 2007.

Curriculum Vitae

Julia Neidhardt is a researcher at the E-commerce Group (Institute for Software Technology and Interactive Systems) at TU Wien, Austria. Her research focuses on modeling complex individual behavior such as traveler preferences in tourism, team collaboration in online games, and peer influence in social networks. The aim of these projects is to utilize novel ways such as gamification and social network analysis to bridge social theories and methods in behavioral sciences and user modeling in data mining, and therefore to develop computational frameworks for individual behavior. She is also working in the area of recommender systems to apply this research.

Julia Neidhardt holds a Master degree in Mathematics from the University of Vienna and has been enrolled in a doctoral program at TU Wien since October 2010. She has developed a course for graduate students on Web Science / Social Network Analysis. She has been teaching this course as well as E-commerce at TU Wien since 2011, and from 2010 to 2012 a course on gender aspects in mathematics at the University of Vienna. In 2013 and 2014 Julia Neidhardt spent several months abroad as a visiting scholar at the *Visual Information Processing for Enhanced Retrieval* (VIPER) research group at the University of Geneva, Switzerland (Short-Term Scientific Missions Grant, COST Action) and at the *Science of Networks in Communities* (SONIC) research group at Northwestern University, USA (Marshall Plan Scholarship, Austrian Marshall Plan Foundation). Julia Neidhardt was a co-organizer and scientific chair of the WWTF Summer School on Digital Humanities that took place at TU Wien in July 2015. She has been supervising several Master theses and has been a reviewer for a number of journals and conferences (including Communications of the ACM, Annals of Tourism Research, Electronic Markets and RecSys conference). She is in the scientific committee of the ENTER conference.

Projects

- Personalisierte Darstellung von Reiseinformationen zur Beschleunigung des Entscheidungsprozesses bei der Reisebuchung (2014 - 2015; FFG Basisprogramm): Analysis of user-generated content from an online travel forum to determine the information needs of the travelers. Methods: text mining, sentiment analysis, machine learning, social network analysis. Role: Julia Neidhardt was responsible for the coordination of the research team and for the design of the analysis. She was also involved in the statistical analysis and the mathematical modeling.

Publications

- [NRW16] Julia Neidhardt, Nataliia Rümmele, and Hannes Werthner. Can we predict your sentiments by listening to your peers? In *Information and Communication Technologies in Tourism 2016*, pages 593–603. Springer International Publishing, 2016
- [DNW16] Amra Delic, Julia Neidhardt, and Hannes Werthner. Are sun lovers nervous? In *ENTER 2016*, volume 7, page 5. e-Review of Tourism Research (eRTR), 2016
- [WASC+15] Hannes Werthner, Aurkene Alzua-Sorzabal, Lorenzo Cantoni, Astrid Dickinger, Ulrike Gretzel, Dietmar Jannach, Julia Neidhardt, Birgit Pröll, Francesco Ricci, Miriam Scaglione, Brigitte Stangl, Oliviero Stock, and Markus Zanker. Future research issues in it and tourism. *Information Technology and Tourism*, 15(1):1–15, 2015
- [SDN15] Dimitris Sacharidis, Amra Delic, and Julia Neidhardt. Learning the role and behavior of users in group decision making. In *Proceedings of the 2nd International Workshop on Decision Making and Recommender Systems*, volume 1533, pages 25–28, 2015
- [NPW15] Julia Neidhardt, Nataliia Pobiedina, and Hannes Werthner. What can we learn from review data? In *ENTER 2015*, volume 6, page 5. e-Review of Tourism Research (eRTR), 2015
- [NHC15] Julia Neidhardt, Yun Huang, and Noshir Contractor. Team vs. team: Success factors in a multiplayer online battle arena game. In *Academy of Management Proceedings*, volume 2015, 2015
- [KPX+15] Worarat Krathu, Christian Pichler, Guohui Xiao, Hannes Werthner, Julia Neidhardt, Marco Zapletal, and Christian Huemer. Inter-organizational success factors: a cause and effect model. *Information Systems and e-Business Management*, 13(3):553–593, 2015
- [NSSW14a] Julia Neidhardt, Rainer Schuster, Leonhard Seyfang, and Hannes Werthner. Eliciting the users' unknown preferences. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 309–312, 2645767, 2014. ACM
- [NSSW14b] Julia Neidhardt, Leonhard Seyfang, Rainer Schuster, and Hannes Werthner. A picture-based approach to recommender systems. *Information Technology and Tourism*, 15(1):49–69, 2014
- [PNCMW13] Nataliia Pobiedina, Julia Neidhardt, Maria del Carmen Calatrava Moreno, and Hannes Werthner. Ranking factors of team success. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1185–1194. International World Wide Web Conferences Steering Committee, 2013

- [PNCM+13] Nataliia Pobiedina, Julia Neidhardt, Maria del Carmen Calatrava Moreno, Laszlo Grad-Gyenge, and Hannes Werthner. On successful team formation: Statistical analysis of a multiplayer online game. In *Business Informatics (CBI), 2013 IEEE 15th Conference on*, pages 55–62. IEEE, 2013

- [PBNW12] Roland Piazzi, Rodolfo Baggio, Julia Neidhardt, and Hannes Werthner. Network analysis of the austrian etourism web. In Matthias Fuchs, Francesco Ricci, and Lorenzo Cantoni, editors, *ENTER 2012, Information and Communication Technologies in Tourism*, pages 356–367. Springer Vienna, 2012

- [FNP+12] Anna Fensel, Julia Neidhardt, Nataliia Pobiedina, Dieter Fensel, and Hannes Werthner. Towards an intelligent framework to understand and feed the web. In *Business Information Systems Workshops*, pages 255–266. Springer, 2012

- [PBNW11] Roland Piazzi, Rodolfo Baggio, Julia Neidhardt, and Hannes Werthner. Destinations and the web: a network analysis view. *Information Technology and Tourism*, 13(3):215–228, 2011

Talks

- Yun Huang and Julia Neidhardt. From networks to space: Constructing metric spaces for social interactions. October 2015. Empirical Investigation of Social Space II, Bonn, Germany; 2015-10-12 – 2015-10-14

- Julia Neidhardt. Computational analyses of network data. Vienna Summer School on Digital Humanities, Vienna; 2015-07-06 – 2015-07-10, 2015

- Julia Neidhardt, Yun Huang, Hannes Werthner, and Noshir Contractor. Conditional random field models as a way to capture peer influence in social networks. June 2015. Sunbelt XXXV, Brighton, UK; 2015-06-23 – 2015-06-28

- Julia Neidhardt. A picture-based approach for travel recommendations. Research Colloquium, School of Computing, DePaul University, Chicago, USA; 2014-10-17, 2014 (Invited Talk)

- Julia Neidhardt, Yun Huang, and Noshir Contractor. Assembly factors influencing the victor in head to head short duration team competitions in a multiplayer online battle arena game. Sunbelt XXXIV, St. Pete Beach, FL, USA; 2014-02-10 – 2014-02-16, 2014

- Julia Neidhardt. Social influence analysis in online travel communities. ENTER Conference 2013 – PhD Workshop, Innsbruck, Austria; 2013-01-22 – 2013-01-25, 2013 (Awarded as 2nd best PhD Workshop Paper)