Analyzing Social Influence in Recommender Systems

PhD THESIS

submitted in partial fulfillment of the requirements for the degree of

Doctor of Technical Sciences

within the

Vienna PhD School of Informatics

by

MSc. Carine Pierrette Mukamakuza
Registration Number 01528430

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dipl.-Ing. Dr.techn. Hannes Werthner
Second advisor: Dimitris Sacharidis, PhD

External reviewers:
Markus Zanker. Free University of Bozen-Bolzano, Italy.
Alessandro Provetti. Birkbeck, University of London, UK.

Vienna, 4th February, 2020

Carine Pierrette Mukamakuza | Hannes Werthner

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

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

Declaration of Authorship

MSc. Carine Pierrette Mukamakuza
Knöllgasse 27/19, 1100 Wien, Austria

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

Vienna, 4th February, 2020

Carine Pierrette Mukamakuza

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

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

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Hannes Werthner, for the patient guidance and for providing his heartfelt support, encouragement, and advice throughout my Ph.D. research studies. I must express my gratitude to my co-supervisor Dimitris Sacharidis, Ph.D. He cared so much about my work and responded to my questions and queries so promptly with a lot of patience. It was an honor to be under their supervision with their exceptional scientific knowledge. Without them this work would have never seen the light today.

My sincere gratitude to the Research Unit of E-commerce, to the Vienna PhD School of Informatics of TU Wien and to the Austrian Agency for international Cooperation in Education and Research (Oead) for awarding me the scholarship to pursue a Ph.D. degree in Austria.

I take this opportunity to thank very much my family; words cannot express how grateful I am to my son for his understanding, love, and patience. Special thanks to Dr. Douglas Mamvura for your encouragement, love, and prayers. They all gave me enough moral support, love, and motivation, and without which I would not come this far.

Many thanks to Clarissa Schmid and Katharina Engel for always being there for me, especially during my peaks and valleys.

All members of EC Research Unit for all their support and their assistance in one way or another that have been of great importance and all people who have been beside me in some way. It's my pleasure to acknowledge them all.

Last but foremost, my sincere thanks to God for everyday protection, care and for giving me the strength to keep going. Without his grace, this achievement would not have been possible.

God bless you.

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

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

Kurzfassung

Ziel dieser Arbeit ist es, den Einfluss sozialer Zusammenhänge auf das Ratingverhalten von Nutzern durch die Untersuchung öffentlich zugänglicher Datensätze zu formalisieren und zu untersuchen. Diese Forschung bietet ein besseres Verständnis für bestimmte Aspekte sozialer Zusammenhänge, die für die Abgabe von Empfehlungen wichtig sind, und trägt somit zur Gestaltung wirksamerer Social Recommender bei. Es ist klar wieso Social Recommender erfolgreich sind: Sie verbessern die Vorhersagegenauigkeit in allen untersuchten Fällen. Die ihnen zugrunde liegenden Annahmen wurden jedoch nicht gründlich untersucht. Wann und wie sollten wir soziale Verbindungen nutzen, um kollaborative Filtertechniken zu verbessern? Gibt es Raum für Verbesserungen bestehender Techniken? Um solche Fragen zu beantworten, müssen wir zunächst die Beziehungen zwischen den Informationsquellen, die einem Social Recommender zur Verfügung stehen – das Ratingverhalten und die sozialen Zusammenhänge – im Detail untersuchen. Wir stellen fest, dass in früheren Studien einige dieser Zusammenhänge unzureichend systematisch untersucht wurden.

Unser methodischer Ansatz berücksichtigt die beiden vorgenannten Sichtweisen, das historische Bewertungsverhalten (V1) und die sozialen Zusammenhänge (V2) der Nutzer. Ziel unserer Studie ist es zu untersuchen, ob Zusammenhänge zwischen diesen beiden Ansichten bestehen. Genauer gesagt definieren wir mehrere Attribute, die wichtige Aspekte jeder Ansicht erfassen, und beobachten dann, ob eine Korrelation zwischen ihnen besteht. Wir unterscheiden drei Arten von Attributen: die, die Benutzer einzeln betreffen und die wir Attribute der Ebene 1 (L1) nennen; diejenigen, die die Beziehungen zwischen Benutzerpaaren quantifizieren, die wir Attribute der Ebene 2 (L2) und Attribute der Ebene 3 (L3) nennen, die Benutzergemeinschaften entsprechen. Für jede Ebene stellen wir Forschungsfragen, die uns helfen, die Zusammenhänge zwischen den beiden Ansichten zu untersuchen und zu verstehen.

In dieser Arbeit wird festgestellt, dass auf allen Ebenen (bei Einzelpersonen, bei Paaren und Freunden sowie innerhalb von Gemeinschaften) signifikante Zusammenhänge zwischen der Bewertung und dem Sozialverhalten (den beiden Ansichten) von Social Recommender bestehen. Wir stellen fest, dass die Stärke der Verbindungen von den untersuchten spezifischen Attributen abhängt und häufig eher gerichtet als bidirektional ist. Darüber hinaus zeigt unsere Analyse bei der Betrachtung von Individuen im Kontext ihres sozialen Umfelds (der Gemeinschaften, denen sie angehören), dass verschiedene soziale

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

Empfehlungsgeber zwar eine vergleichbare Wirksamkeit aufweisen, sich jedoch in ihren Auswirkungen auf die Präferenzen von Individuen unterscheiden. Insgesamt liefert diese Arbeit einige konkrete Beiträge zum besseren Verständnis der Auswirkungen sozialer Zusammenhänge auf das Ratingverhalten von Nutzern in sozialen Empfehlungssystemen. Darüber hinaus wird in dieser Arbeit ein Social Recommender vorgeschlagen, der genauso effektiv ist wie vorhandene Techniken und die Benutzer fair behandelt.

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

Abstract

The aim of this thesis is to formalize and investigate the degree of impact that social connections have to the rating behavior of users, by studying publicly available datasets. This research provides a better understanding of specific aspects of social connections that are important when making recommendations, and thus contributes towards designing more effective social recommenders. It is clear that social recommenders are successful: they improve the prediction accuracy in all cases examined. However, the assumptions they are based on, have not been thoroughly studied. When and how should we use social connections to augment collaborative filtering techniques? Is there room for improvement in existing techniques? To answer such questions, we must first examine in detail the relationships between the two sources of information available to a social recommender, the ratings behavior and the social connections. We note that although previous work has investigated some of these relationships, it has done so in a non-systematic way.

Our methodological approach considers the aforementioned two views, the historical rating behavior (V1), and the social connections (V2) of users. The goal of our study is to examine whether connections between these two views exist. More concretely, we define several attributes capturing important aspects of each view, and then observe whether there is a correlation between them. We discern three types of attributes: those that concern users individually, which we call level 1 (L1) attributes; those that quantify relations between pairs of users, which we call level 2 (L2) attributes and level 3 (L3) those that correspond to user communities. For each level, we pose research questions that help us explore and understand the connections between the two views.

This thesis finds that there exist significant connections between the rating and the social behavior (the two views) in social recommenders at all levels (among individuals, among pairs and friends, and within communities). It also finds that the strength of the connections depends on the specific attributes examined, and is often directional, rather than bi-directional. Moreover, when looking at individuals in the context of their social circle (the communities they belong to), our analysis shows that although various social recommenders have comparable effectiveness, they differ in the impact they have on individuals' preferences. Overall, this thesis makes several concrete contributions towards a better understanding of the impact of social connections to the rating behavior of users in social recommender systems. Furthermore, this work proposes a social recommender that is as effective as existing techniques and treats users rather fairly.

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

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

Contents

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

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

CHAPTER 1

Introduction

1.1 Overview

Information overload is a characteristic of our society today. This has made decision making more complicated than ever before. As an example, choosing which suit to wear and with which pair of shoes and jewellery to match it with becomes a challenge to some people. So is choosing which restaurant to go to for the day's meal further complicated by a variety of dishes on the same menu. In the entertainment industry, one has to choose whether to read a book and if so which book or watch a movie and if yes, which one? Subsequently, people have resorted to social connections to seek for advice and expert opinion regarding what choice to make under different circumstances. Recommender systems have come in handy as far as overcoming such challenges is concerned. They attempt to guide people into making decisions based on their preferences, their personality and by mimicking the choices of people similar to them. The key idea of Social Recommender Systems (SRS) is to enhance recommendations by drawing information from the social context of the user. The underlying assumption is that for a particular item, the decision making process of a user not only depends on their individual preferences, but also on interpersonal influence from their social connections. For instance, influential people may strongly affect the decision making of a person and thus, the structure of a social network is important in trying to understand the social effect and the extent of its impact. Therefore, this study focuses on collaborative filtering (CF) based social recommenders that draw information from two components. The first is the rating behavior represented by the ratings matrix where each existing entry corresponds to the rating given to an item by the user. The second component is the social connections cconveyed by the social adjacency matrix, where entries portray the friendship strength between users. Social recommenders predict ratings using these two matrices under the assumption that a user's behavior is influenced by their social connections.

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

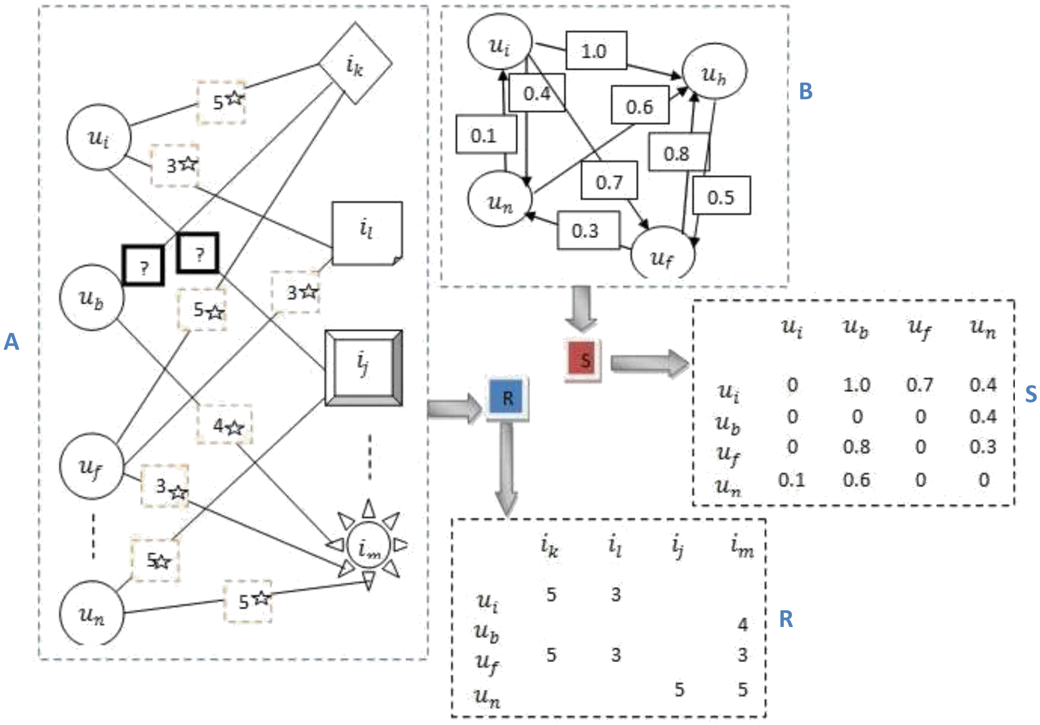


Figure 1.1: An example of user-item rating behavior and user-user social connections.

An example of a social recommender is shown in Figure 1.1, which depicts the rating behavior of users denoted as u_i, on items denoted as i_j on the left (A) and the social connections among users on the right (B). The former is captured by the rating matrix R, where a non empty entry R_{ij} corresponds to the rating given by user u_i on item i_j. The latter is conveyed by the social adjacency matrix S, where entries portray the friendship strength between users. Social recommenders draw upon information from both matrices to predict ratings.

Despite the fact that SRS are relatively recent, they have become an active area of research over the past few years. This is partly because existing approaches are plagued with a number of weaknesses such as making explicit assumptions about the impact of social ties that they never validate[Muk17, MSW18]. They also fail to take into account the structure (local and global) of the social network and the magnitude of the impact it has on the rating behavior. Our main aim is the formulation and (statistical) analysis of the impact that social connections have in rating behavior at different levels. Can we predict how users rate items, and to what extent, purely by observing their position in the social network and vice versa? An additional contribution is the theoretical evaluation of the assumptions made by state of the art social recommenders, and whether they hold in various domains. Ultimately, we would have a better understanding of what aspects of

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

social connections exactly affect rating behavior. This will bring us initial ideas towards a more realistic model for social recommendations, based on observed and quantifiable types of social influence.

1.2 Research Motivation

The idea of social recommendation is to improve recommender systems by incorporating social ties, social contextual, information[JCL+12] which can be derived from links on social networks and can be explained as follows: Given an item, the behavior of user depends on individual preference to understand whether the user likes it or not, and interpersonal influence to tell whether the user has tight relationships with the item advocates (e.g. those that like the item) or not; influential people may influence the behavior of a friend, a group or a community. A social based recommender system is a way of considering social network information to improve recommender systems. The network structure is very important while trying to understand the Social influence.

There is a common assumption in social recommenders: if two people are socially connected, then they must have similar preferences. This assumption is adopted by proposals to a different extent. Some of the proposed methods, e.g., [MYLK08, MKL09, JE10], go to the extreme, as they explicitly mandate that two friends should have similar preferences (user features). This approach completely ignores the fact that the degree of influence/homophily may actually vary among friends.

More recent methods based on social regularization, e.g., [MZL+11, LWTM15, ZYKL17], acknowledge that not all pairs of friends should be treated equally. Instead, they force two friends to have similar features to the degree that their observed rating behavior is similar. At first, this may seem like a more realistic model, but upon a more detailed inspection, we find that it defeats the purpose of using a social recommender. If two friends exhibit very similar rating behavior, then any good CF model that is agnostic to the two users' social connections, should be able to understand this relationship on its own, and assign similar features to these users anyway. All social regularization does is to make this even more explicit for the underlying model, asking it to ensure that similar friends have similar features. So in this case, social recommenders do not do anything different than plain old collaborative filtering.

Now consider the case of two friends, between which at least one is a cold-start user so that their observed rating behavior is not similar — at least so far. In this setting, a social regularization-based recommender would mandate that these two friends should not be assigned similar features, again much like basic collaborative filtering would. However, this ignores the possibility of social influence between these friends, which is exactly the premise behind social recommendation: when there is little information in the ratings matrix to work with, augment it with social connections.

It is clear that social recommenders are successful: they improve the prediction accuracy in all cases examined. However, the assumptions they are based on, have not been

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

thoroughly studied. When and how should we use social connections to augment collaborative filtering? Is there room for improving existing techniques? To answer such questions, we must first examine in detail the relationships between the two sources of information available, the ratings (what we call view V1) and the social connections (view V2). We note that although previous work, e.g., [SLA10], has investigated some of these relationships, it has done so in a non-systematic way. We believe that our research approach can bring novel insight that can help design more effective social recommenders.

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

1.3 Problem Statement

The thesis investigates whether there exists a relationship between social connections and rating behavior. For this purpose, we study publicly available datasets, which are commonly used in literature, containing both user-item ratings and user-user connections. We employ collaborative filtering techniques to associate users based on their behavior, and network analysis methods to associate users based on their connections, and examine whether correlations exist at different levels. In particular, we consider three levels: individual users (L1), pairwise (L2), and community (L3) and pose research questions that seek correlations between social connections and rating behavior.

1.3.1 Overall appoach

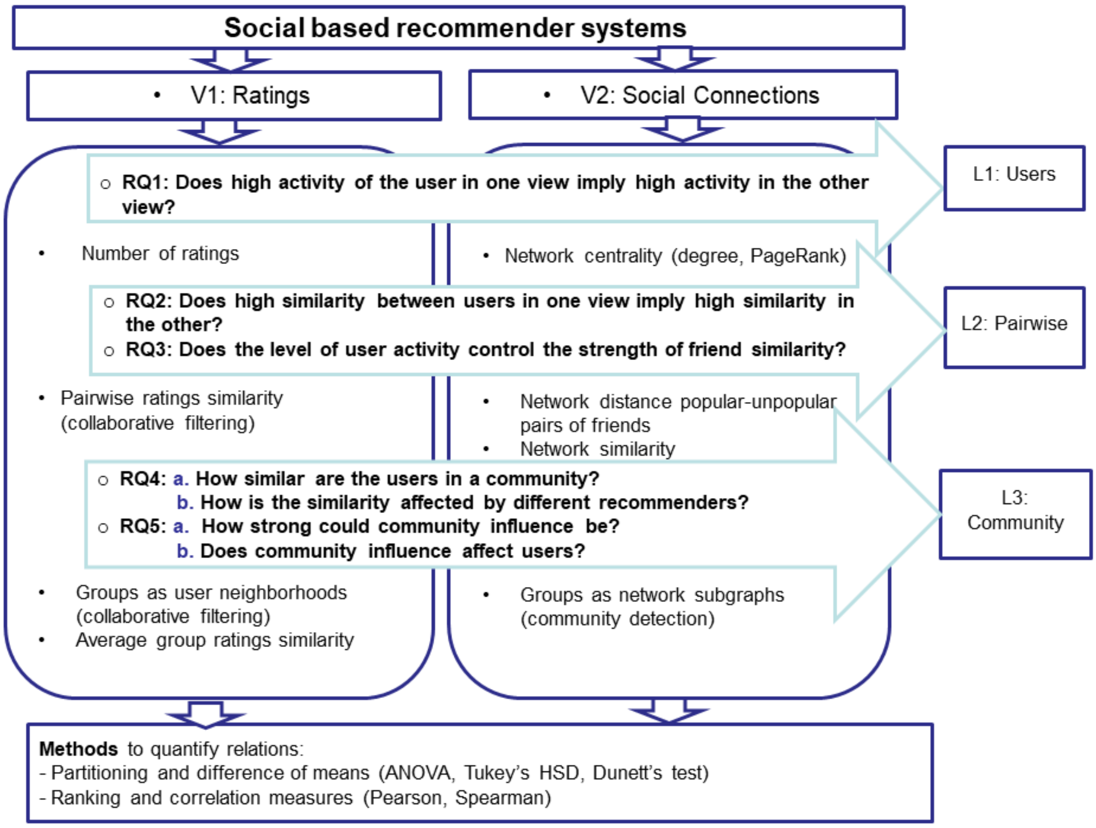


Figure 1.2: An illustration of our methodology for studying relationships between rating and social behavior in social recommender systems.

Our methodological approach is illustrated in Figure 1.2. We consider two *views*, the historical rating behavior (V1), and the social connections (V2) of users, corresponding to the two sources of information available to a social recommender. The goal of our

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

study is to examine whether connections between these two views exist. More concretely, we define several *attributes* capturing important aspects of each view, and then observe whether there is a correlation between their value distributions. We discern three types of attributes: those that concern users individually, which we call level 1 (L1) attributes; those that quantify relations between pairs of users, which we call level 2 (L2) attributes and level 3 (L3) those that correspond to groups and communities. At each level, we consider several attributes, from both views, describing the object of study (users or pairs of users or communities), and we seek to quantify correlations between attributes of different views.

1.3.2 Research questions

What are the characteristics of social recommender systems datasets?

Many researchers have been using the following datasets to do their experiments to compare the qualities of recommendation: Epinions, Flixster, Douban, Dianping, Yelp, Renren, TencentWeibo, etc. Most of these datasets are composed of users, social links (trust, friendship, communities), Items, ratings, tweets, cold-start users both rating cold-start users and social cold-start users, etc. Another important issue is the structure of these datasets. Social networks generate varied datasets and subsequent properties. Some methods therefore, cannot be applied to some datasets, the dataset require some level of understanding before further processing. The idea is to obtain basic statistics and insights of the general description of the dataset, in order to guide our research decisions made later on.

RQ0: Which Social Recommender System performs best?

RQ1: Does high activity of the user in one view imply high activity in the other view?

The first research question is based on the first level (level of users), where the challenges are the network analysis while getting users' structure in the network in order to identify important users, and extraction of ratings for each user in the rating matrix, then compare both our outputs in both sides (in the network and the ratings).

For example, *Are heavy raters popular?* or is there a connection between *heavy raters*, who have made a large number of ratings, and *popular users*, who have acquired many social connections in the system? To answer this question we go both ways, looking whether heavy users are popular and vice versa. Specifically, we employ techniques from social network analysis to determine different interpretations of "*popularity*" based on network centrality. On the other hand, the "*heaviness*" of a user has a single interpretation, the number of her ratings.

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

RQ2: Does high similarity between users in one view imply high similarity in the other?

The second research question is based on the second level (pairwise level), where the challenges are computing ratings similarity between users from the rating matrix, computing friendship strength between users in the Network, then check from the social network (SN) if the friendship strength results give a reasonable relation with the similarity results.

For example, the sub-question in this case is *Do Friends have similar ratings?*

RQ3: Does the level of user activity control the strength of friend similarity?

The third research question is based on the second level (pairwise level), where the challenges are to find user's position or importance in the SN (descriptive analysis) then classify them into categories; High centrality users (H), Low centrality users (L) and Pairs of friends (H-H; H-L; L-L) then answer the following question: Is the correlation to their pairs getting stronger?

If we know individual aspects about users, e.g., their level of activity in a personalization system, can we infer a pairwise relationship, e.g., the similarity of their observed activities, between friends? For example, *Do popular users influence more?*

RQ4: How similar are the users in a community? How is overall similarity affected bydifferent methods?

This is about the third level (communities). If we consider the rating neighborhood of a user in a Matrix factorization manner, do we see strong social connections among these users? One way to answer this is to compute for every user the similarity of ratings in her neighborhood, and how many similar neighbors are in his/her community. Then look across all those neighbors if the rating similarity is significant.

RQ5: How strong is the community influence? Does community influence affect users?

This question is also on the third level. Suppose we have a way to identify communities by different community detection methods in the social network and also quantify their strength based on community influence. Then we can compute the ratings similarity of each community by using different social regularization based recommenders. Does the community strength correlate with ratings similarity and how is the community strength affected by different methods?

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

1.4 Research Methodology

In this section, we position our contributions based on Design Science Research (DSR) knowlege contribution framework. DSR knowlege contribution framework was introduced by [GH13] and the main goal is to classify the research contributions in four categories based on 2x2 provided matrix. The four quadrants of the matrix are introduced briefly:

- Improvement: As the name of the quadrant sounds, the quadrant stand for *new solutions for known problems* here the contribution is represented by the work done to bring new solution to existing/known problems that have been solved poorly or inefficiently. The key is to come up with a solution that gives an improved products, services, ideas.
- Invention: this quadrant is for *new solutions for new problems*, the contribution is considered to be categorized in this quadrant once the problem is fresh in the field and the solution as well is invented.
- Routine Design: This quadrant represents *known solutions for known problems*, in general this quadrant is not normally thought of as contribution but in some cases this type of work leads to some interesting and amazing findings. Routine design actually has a lot to do with existing solutions that are used to solve known problems as well.
- Exaptation: This quadrant stands for *known solutions extended to new problems*. The contribution in this quadrant is to initiate the existing solutions and extend their usability to some new application area, as in some new field solution are still very few, so for example one existing model can be used in many different fields.

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

Figure 1.3 illustrates the matching position of methods and levels. Methods are explained in details in the following subsections.

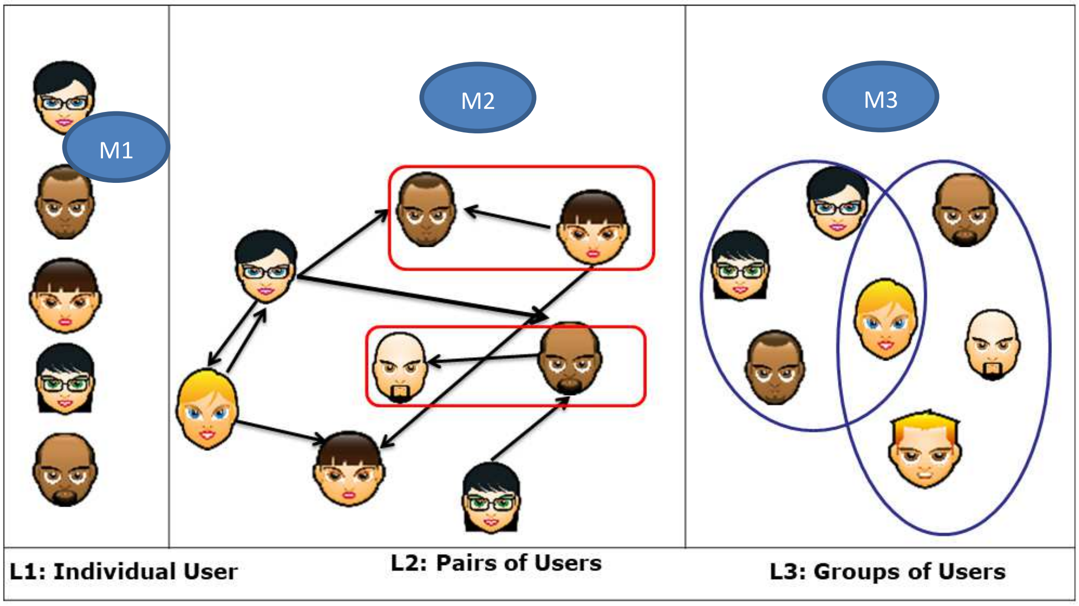


Figure 1.3: Methods and Levels

Figure 1.4 represents DSR Knowledge Contribution Framework with reference to our contributions.

1.4.1 M1

M1 level encompasses methods used at individual level and takes into consideration the uniqueness of each user's properties. For example, her importance and her popularity based on her connectivity position or location in the network at the V2. At the V1, rating frequency is taken into account to observe the rating behaviour of every user individually. For this matter, the network centrality (degree centrality) and PageRank are used to determine the user importance and popularity in the network which are known as basic network analysis measures at microscopic level. To study the relationship between V1 and V2, two methods are used: The first is Partitioning (Partition users) which is based on quantity of one aspect (e.g. social connections) and examines how another quantity of the other aspect (e.g. rating behavior) varies across partitions. Users are grouped into different groups based on their activity level by using one chosen attribute, for example degree centrality if social network aspect is considered. In case rating behaviour aspect is considered the partitioning is based on rating attribute. The second is Ranking (Rank users) which is based on two quantities, one for each aspect, and then measure the ranking correlation using standard techniques.

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

1.4.2 M2

M2 level encompasses methods presented at Pairwise level. Pairwise level is the level where pairs are playing the main role to be able to learn the impact of the connections or ties on the target user. The first analysis brings up how similar or dissimilar connected users are and also the distance between a user and her ties is deduced in order to extract the impact of the distance on a user and her peers. At Pairwise level we also studied the impact of having popular friends and how influential they can be. Macroscopic level network analysis is taken into account. Partitioning and Ranking methods together with statistical metric tool such as ANOVA are used.

1.4.3 M3

M3 level encompasses all methods used at Community level, the level where we consider a group of users and study how they can influence each others in terms of rating. Two different algorithms are used to determine user communities called influencer based communities. This takes into account the users' popularity by considering the level of popularity of each user to her ties. Another algorithm is modularity based community which is greedy modularity community. This involves making communities by first setting each user as a community then maximize the modularity until the pairs are formed.

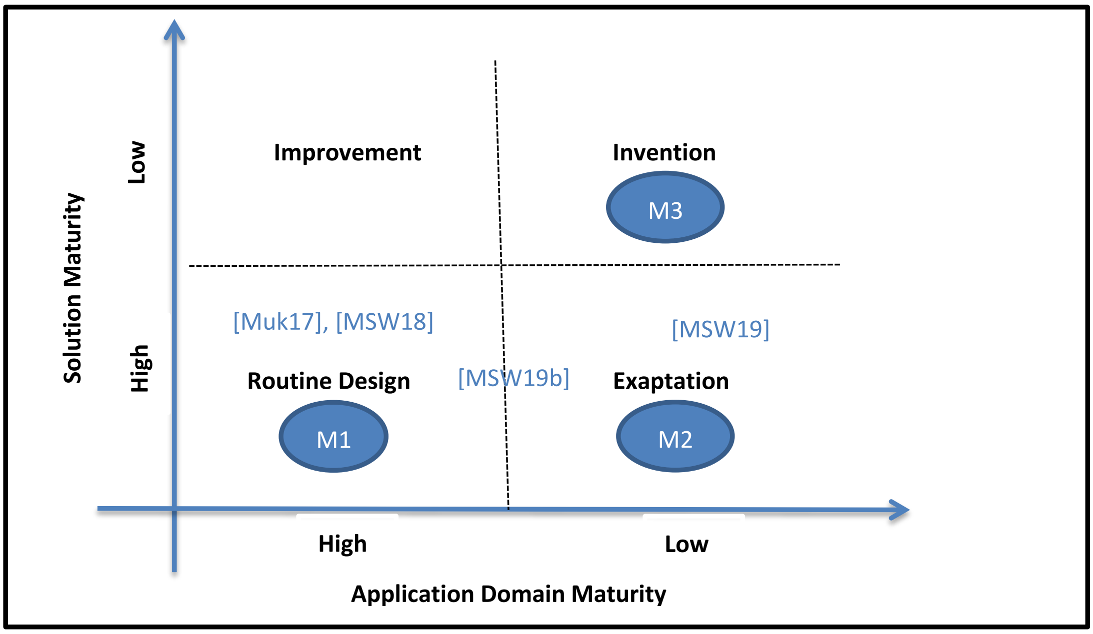


Figure 1.4: DSR Knowledge Contribution Framework

The first result we achieve is to statistically analyse the effect of social ties in ratings, where we observe the behaviour of users when giving ratings and the impact of the

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

ties they set up, support, end, etc. The influence of preference for interaction with people/group that are similar to you and the way network structure may also affect the different ties behaviour according to user's position.

Another contribution is the theoretical evaluation of the assumption made by state of the art of social recommenders, investigates whether the assumptions hold in various datasets. For example, in RSR its believed that our friend's recommendations will have a big impact on our choices and decisions because we believe in the tastes and suggestions of our friends.

At the end we will reach a better understanding of what exactly impacts rating behaviour on different levels; individually, in pairs and in a community which brought us the initial ideas towards a more realistic model for social recommendations that is based on actual data and on observed types of social influence.

1.5 Outline

The remaining part of this dissertation is organized as follows:

- *Chapter 2*: Reviews the state of the art of Social based recommender systems. It starts by introducing social network analysis methods, RS techniques and presenting statistical metrics used in this dissertation. The goal of this Chapter is to build a mental picture of existing solutions to problems stated in Section 1 and also exposes the gap that is yet to be filled by existing solutions, which we aim to bridge. As an important aspect in this field, social based recommender systems are also discussed.
- *Chapter 3*: Gives details of methods used in order to answer the research questions and give insight on how we have chosen to answer them based on three different levels with three different methods (M1, M2 and M3) which are explained in details as well as in the sections of this chapter.
- *Chapter 4*: Gives details of the study done on the first level (Individual level) which comprises the first research question (RQ1) as it is shown in Figure 1.2. Experimental results and discussions are also given.
- *Chapter 5*: A detailed experimental analysis conducted on the second level (pairwise level) is given in this chapter. This level's experiment is based on two research questions (RQ2 & RQ3), experimental results and discussions are also given. Note that in this chapter, we also present our models and evaluate them.
- *Chapter 6*: In this chapter, we answer the questions that we formulated on the third level (level of communities), we extract communities in the network using network community extraction methods and present the State of the Art evaluation. We use different matrices as regularizers, the Matrix Factorization (MF) model is the baseline. In the end, we study the relationship of users in the same community and their rating similarity based on different methods.

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

Concluding remarks, lessons learned and future discussions are found in Chapter 7.

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

CHAPTER 2

Related Work

This chapter gives a background picture and review of existing methods in Recommender system, Social network analysis and Social based recommender systems. In addition to that we also point at relevant applications of recommender system and social recommender systems.

2.1 Recommender Systems

2.1.1 History Of Recommender Systems

The very first recommender system came in the late 70's which is fairly early in the history of computers. The name of the recommender was Grundy and it was a system used for a library with the main goal of suggesting novels to people who were first organized into different stereotypes. It was pretty impressive at the time it was designed as it incorporated the personalities and goals of all the distinct users before making recommendations. About twenty years later we saw the rise of collaborative filtering which came as a solution to the huge overload in data. One of the first systems that used collaborative filtering was Tapestry which allowed users to search for items in an information domain based on the opinions of other users. Tapestry was followed by Grouplens which introduced automated collaborative filtering recommendation systems. Grouplens' main goal was to suggest interesting 'Usenet articles' by finding similar opinions between different users. The idea is that the active user can express whether they like a Usenet article or not in which case the system predicts and recommends articles that may probably be liked basing on people with whom the user has a shared taste. This is the nearest neighbor method which method I adopt in this dissertation. Collaborative filtering became widely known leading to increased interest in machine learning and data mining generally. Various recommender systems were introduced such as Bellcore Movie Recommender and Ringo Music Recommender. Furthermore,

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

during this time recommender systems were also used in marketing and became really useful for increasing sales as well as generally reducing the the time and effort customers used in search for items to consume. Their shopping experience was also subsequently improved. Not long after, Amazon was created. Amazon uses a collaborative and content based filtering method along with what the user is browsing at the time, to make its recommendations. Significant growth in the study of recommender systems was realised in 2006, when Netflix launched their Netflix prize competition to improve their movie recommendation algorithm. Today, Netflix is considered to have one of the most advanced hybrid recommender systems.

What Exactly Is A Recommender System?

Simply speaking, a recommender system is one that has the ability to collect and present to individual users a variety of information (in a general sense) in line with their field of interest. This, specifically is referred to as a recommendation and in simple terms, it is the collection and presentation of useful information to a specific user. This definition has been expanded by researchers with examples such as [Bur02], where they state that a recommender system is "any system that produces individualized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options". [AT05] made it even more formal by stating that "the recommendation problem can be formulated as follows: Let U be the set of all users and let I be the set of all possible items that can be recommended. Let u be a utility function that measures the usefulness of item i to user u, that is , $u : U \times I \rightarrow R$, where R is a totally ordered set (for example, non-negative integers or real numbers within a certain range). Then for each user $u \in U$ we want to choose such items $i^{'} \in I$ that maximizes the user's utility". We can clearly see from the above definition that the goal of a recommender system is to generate and recommend to users items with the best correlation and not just to predict the correlations between the users and all the different items. Against this background two significant facts stand out regarding recommender systems. First, personalization is an important part of a recommender system given that its main focus is to ensure the recommendation of specific products and services to a particular user and not represent group consensus for all users. Secondly, the recommender system should have some basic information about the user with which to make suitable and/or appropriate recommendations. This is made easier if the user has a good number of options, including preferred items known in advance and not randomly generated.

Recommender System Today

Recommender systems have become an important effortless decision making in our daily life. RS have made life much easier for browsers, this comes in handy when they have limited time and/or patience and are not sure of what they are looking for. Here we can specifically point at YouTubers' followers and fans, as well as in variety of areas such as Netflix for music and movies entertainment, for shoppers at Amazon and many other

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

online stores. For online stores they play the part of the sales person, one that knows all the data or number of products the online service offers and knows with a given level of certainty what you like and do not like. Professionals also enjoy RS services when it comes to seeking important people to link with on LinkedIn or Twitter. In tourism domain RS is used a lot in many different ways, e.g. we can list the group recommendation while deciding which place to visit as a group of people (friends, co-workers). In our normal social life, recommender systems are utilized to help us rediscover our long lost friends (Facebook).

So, recommendations help solve the problem of discovery by providing top picks for you, suggestions in the style of "If you like this, you will also like that" and "if you buy this, you will need that". They do that with the use of huge databases including, among other things, what the users browsed, what they bought, what and when they clicked and what they rated.

2.1.2 Recommender Systems Major challenges

Researchers in the field of recommender systems face several challenges. Here we mention only the major ones.

Data sparsity

This problem rises because the pool of available items is often extremely large compared to very limited number of items that users usually rate. The generated inadequate data simply called sparse user-item rating matrix is generated from non-classified items. A large amount of unknown entries in the system with such sparse data makes identifying similar users based on the number of ratings provided harder. The quality of recommendations is negatively impacted by this phenomena of the lack of enough feedback needed to predict. Researchers attempt to address data sparsity by using many different methods including the matrix factorization model [New05]. Therefore, an effective recommender algorithm must take the data sparsity issue into account [HCZ04].

Scalability

While the data is mostly sparse, for major sites it includes millions of users and items. As such, it is essential to consider the computational cost issues and search for recommender algorithms that are either less demanding or easy to use in parallel (or both). Another possible solution is based on using incremental versions of the algorithms where, as the data grows, recommendations are not recomputed globally (using the whole data) but incrementally (by slightly adjusting previous recommendations according to the data coming in) [SKKR02, JLZZ09]. This incremental approach is similar to perturbation techniques that are widely used in physics and mathematics [SPUP02].

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

Cold start

Due to new users or new items that enter the system, the problem of Cold start is heavily addressed in the literature. There is usually lack of sufficient information about new users' preferences which makes it almost impossible to produce recommendations for them. In such cases the ideal solution is hybrid recommender techniques because of their combination of content and collaborative techniques [ÇM19]. Authors in [MCLZ18] extract hidden features from item's representation by using probabilistic model and they generated accurate pseudo ratings from extracted hidden features, even in cold-start case when there is a small number or no ratings are provided at all. Sometimes some additional basic information is required from users such as their location, age, gender [KC08, OJ06]. Another way is by identifying individual users in different web services. For example, Baifendian [ZXNL15] developed a technique that could track individual users' activities in several e-commerce sites, so that for a cold-start user in site A, we could make recommendation according to her records in sites B, C, D and so on.

Black and Grey Sheep

Grey sheep occurs when one user's opinion is not in agreement or disagreement with the group of users of the system. This user will hardly receive a worthwhile recommendation. *White sheep* on the other hand is the user who classifies the items in the same way as other users while the black sheep user rates the items as extremes (very good or very bad) and thus has few or no group of users to relate with [GIL+09].

Beyond Accuracy

The very important task of recommendation is to make sure the user is satisfied by the recommended items, most of the time it is effective to recommend highly rated and popular items. The problem with such recommendation though is that popular items do not stay long in the store they quickly run out of stock and by taking into account that different users have different tastes, popular object to one use might be unpopular to another. Hence, a good list of recommended items/objects should contain also less obvious items that are unlikely to be reached by the users themselves [dMP97]. Approaches to this problem include direct enhancement of the recommendation list's diversity [EH05, HZ11, ZKL+10] and the use of hybrid recommendation methods [MBBW07]. In the past few years, many researchers come to realize that accuracy as evaluation metric alone is not enough to identify the effectiveness of a recommendation functionalities. Diversity and novelty metrics are the most important key qualities beyond accuracy in real recommendation scenarios when it comes to measure the utility. The relation between both is that the diversity specifies how the set of items is diverse which means how different those items are to each other; the diverse set of items implies novelty as each item is novel with respect to each other.

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

Interpretability

From the browsers' side interpretability can be very helpful to express why the specific recommendation is offered to her. Youtube provides a solution to help users with such concerns while recommending videos a historical link for the user is added to show the user what triggered the recommendation. Interpretability is also very important for the modelling side as it helps developers to understand how the system function. Traditional collaborative approaches are easy to interpret their results, in a manner that the approach methodology for a recommendation could be written down into a simple human understandable sentence. Modern Collaborative filtering using latent space models is very hard to understand compared to content-based approaches. From the modeling side, Content-based approaches are easy to interpret, while collaborative filtering models are harder to understand. One can cluster the items or users based their original feature space or latent space (matrix factorization and deep learning), and check whether the objects from the same cluster share similar characteristics.

Vulnerability to attacks

Due to their importance in e-commerce applications, recommender systems are likely targets of malicious attackers trying to unjustly promote or inhibit some items [MBBW07]. There is a wide scale of tools preventing this kind of behavior which ranges from blocking the malicious evaluations from entering the system to sophisticated resistant recommendation techniques [KKLP06]. However, this is not an easy task since the strategies of attackers also get more and more advanced in the same way development of preventing tools does. For example, Burke et al. [BOH15] introduced eight attacking strategies which are further divided into four classes: basic attack, low-acknowledge attack, nuke attack and informed attack.

The value of time

While real users have interest in widely diverse time scales (for example, short term interests related to a planned trip and long term interests related to the place of living or political preferences), most recommendation algorithms neglect the time stamps of evaluations. Ongoing research currently revolves around the value of old opinions for example whether they should decay with time alongside typical temporary patterns in user evaluations and item relevance [MH05, RKTY10].

Evaluation of recommendations

While we have plenty of distinct metrics, how to choose the ones that best correspond to a given situation and task is still an open question. Comparisons of different recommender algorithms are also problematic because different algorithms may simply solve different tasks. Finally, the overall user experience with a given recommendation system, which includes user's satisfaction with the recommendations and user's trust in the system is

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

difficult to measure in online evaluation. Empirical user studies thus still represent a welcome source of feedback on recommender systems.

User interface

It has been shown that to facilitate users' acceptance of recommendations, the recommendations need to be transparent [TW02, HZC07]: The user's appreciation level is high when it is clear why a particular item has been recommended to them. Another issue is that since the list of potentially interesting items may be very long, there is need for recommendations to be made in a simple way allowing for easy navigation. There should also be a way to browse through different recommendations which are often obtained by distinct approaches. Besides the above long-standing challenges, many novel issues appear recently. Thanks to the development of methodology in related branches of science, especially the new tools in network analysis, scientists started to consider the effects of network structure on recommendation and how to make use of known structural features to improve recommendation. For example, Huang et al. [MMC17] analyzed the consumer-product networks and proposed an improved recommendation algorithm preferring edges that enhance the local clustering property, and Sahebi et al. [SQBB10] designed an improved algorithm making use of the community structure. Lastly, intelligent recommender systems should take into account the different behavioral patterns of different people.

Preference elicitation

Online services have introduced explicit and implicit elicitation as most used methods for new user preference elicitation. In some systems like Netflix, for example, users rate items on the scale of one to five stars in explicit elicitation[CHT15, KWG$^+$12]. In implicit elicitation, preferences are obtained from purchase and browsing history, search patterns and even mouse movements[HKV08] such as Amazon, Youtube and Facebook. A Problem occurs when a user got into preference conflicts, while searching for an item on Amazon that has a low price and still a good quality, for example, mini-projector Qumi with high resolution and then get the reply "nothing found" because the user is only allowed to enter preferences only one at a time. The recommendation accuracy has been improved by many researchers by combining both explicit and implicit elicitation. However, implicit feedback is considered useless for the system that focuses on personalization. Also many studies have not done enough investigations on users' experiences and behaviour of implicit and explicit elicitation.

2.1.3 Recommender Systems Techniques

Recommender Systems (RS) are systems that are able to suggest or provide items to the target user. In RS, the user rating matrix is composed of items and users sets. A set of items $I = i_1, ...i_n$ and a set of users $U = u_1, ...u_m$. The ratings matrix $R = [R_{u,i}]n \times m$ where n represents the number of items and m number of users. The most commonly

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

used method traditionally in prediction is collaborative filtering (CF). Under the CF approach (Memory based and Model based), users with similar rating patterns are taken into account [JCS04]. The ratings of these similar users on the target item i is aggregated in CF to compute the predicted rating for user u. Memory based prediction ratings of user u based on the ratings of item by a set of the users whose similarity level is closer to the target user. Model based method makes prediction by learning parameters.

2.1.4 Collaborative Filtering

As explained in the previous section, CF includes two major classes; the first one is Memory based which uses two different approaches. User-based technique is one of them which uses the assumptions: if users had similar tastes in the past, they are most likely to have the same tastes in the future and users prefer to remain constant and stable over the time. This approach uses user u_i profile (rating vector) and other user profile for example u_f to compute the similarity value and predict the rating of the target user. Item-based uses the target user's profile to compute items similarity value, for example similarity between item i_j and i_k. The rating scale is mostly numbers from 1 to 5, where 1 is strongly disliked and 5 strongly liked. In memory based method, user-based and item-based [SKKR01] recommendations, there are two steps, the neighbourhood formation and prediction steps. In user-based, the formation of neighbourhood phase is computed this way: Given rating matrix R computer similarity between target user u_i and neighbour user u_f:

$$sim(u_i, u_f) = \frac{\sum_{i_j \in I}(r_{u_i i_j} - \bar{r}_{u_i}).(r_{u_f i_j} - \bar{r}_{u_f})}{\sqrt{\sum_{i_j \in I}(r_{u_i i_j} - \bar{r}_{u_i})^2}.\sqrt{\sum_{i_j \in I}(r_{u_f i_j} - \bar{r}_{u_f})^2}}$$

Where r_{ij} is the rating given to item i_j by user u_i and $\bar{r}_i$ is the average of all ratings given by u_i. To predict rating, users with the highest similarity value are considered.

The rating prediction phase for the item i_j for target user u_i is given by the formula:

$$p(u_i, i_j) = \bar{r}_i + \frac{\sum_{u_f \in U_{u_i}} Sim(u_i, u_f) \times (r_{u_f i_j} - \bar{r}_{u_f})}{\sum_{u_f \in U_{u_i}} |Sim(u_i, u_f)|}$$

Where U_{u_i} is user u_i's neighborhood (the set of top-h similar users). User based predicts ratings based on users. Item based recommendation predicts ratings based on the items, as mentioned above it also has two phases the neighbourhood formation and the prediction. Item based recommendation [Bur12].The similarity of item i_j and i_k is computed as follow:

$$sim(i_j, i_k) = \frac{\sum_{u_i \in U}(r_{u_i i_j} - \bar{r}_{u_i}).(r_{u_i i_k} - \bar{r}_{u_i})}{\sqrt{\sum_{u_i \in U}(r_{u_i i_j} - \bar{r}_{u_i})^2}.\sqrt{\sum_{u_i \in U}(r_{u_i i_k} - \bar{r}_{u_i})^2}}$$

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

The prediction of item i_k to target user u_i is given by the cosine similarity:

$$p(u_i, i_k) = \frac{\sum_{i_j \in I_{u_i}} r_{u_i,i_j} \times Sim(i_j, i_k)}{\sum_{i_j \in I_{u_i}} Sim(i_j, i_k)}$$

Where i_j belong to rated items I_{u_i} of u_i.

Model based recommendation gives prediction by learning parameters and building models. The commonly used methods are matrix factorization and Probabilistic recommendation approaches. Matrix Factorization (MF) is explained in the section below.

2.1.5 Basic Matrix Factorization (MF)

Basic Matrix Factorization considers $m \times n$ rating matrix R which describes m users numerical ratings on n items[MZL+11].A low rank matrix factorization approach tries to approximate the rating matrix R by multiplication of l-rank factors [SM07]. R represents the rating matrix, m users, n items. In matrix factorization, rating matrix R is decomposed in two matrices: U User-aspect matrix and V Item-aspect matrix. l represent latent factors/features. A rating r_{ij} can be given by a dot product of vector U_i and item V_j:

$$R \approx U^T V$$

Where

$$U = [U_1, U_2, U_3, ...U_m]$$

$$V = [V_1, V_2, V_3, ...V_n]$$

,$U \in R^{l \times m}$ and $V \in R^{l \times n}$ with $l < min(m, n)$. As in the real life, each user only rates a very small part of items, the matrix R is usually extremely sparse[RS05]. The goal is to minimize reconstruction error:

$$\frac{1}{2} \|R - U^T V\|_F^2$$

where $\|.\|_F^2$ denotes the Frobenius norm. However, due to the reason that R contains a large number of missing values, we only need to factorize the observed ratings in matrix R. Goodness of fit is used to decrease the prediction errors:

$$\min_{U,V} \sum_{i=1}^{m} \sum_{j=1}^{n} I_{ij}(R_{ij} - U_i^T V_j)^2$$

Where I_{ij} is the indicator function that is equal to 1 if user u_i rated item v_j and equal to 0 otherwise. Two regularization terms are added to alleviate overfitting:

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

$$\min_{U,V} \sum_{i=1}^{m} \sum_{j=1}^{n} I_{ij}(R_{ij} - U_i^T V_j)^2 + \frac{\lambda_U}{2} \|U\|_F^2 + \frac{\lambda_V}{2} \|V\|_F^2$$

Where $\lambda_U, \lambda_V > 0$. The optimization problem minimizes the sum-of squared-errors objective function. Gradient based approaches can be applied to find a local minimum. The above algorithm is perhaps one of the most popular methods in collaborative filtering.

2.1.6 Content based filtering (CBF)

Content-Based Filtering (cognitive filtering), recommends items based on the assumption that users who had the same taste of items with certain attributes in the past, will most likely, like the same kind of items in the future. Content based filtering works in a way that the item content is represented as a set of terms; typically it makes use of item features, simply by comparing the words that occur in a document with the user profiles, this is made up of the same set of terms that is generated by content search history analysis of the user. CBF has advantages; the first one, it does not need other user's data to make a recommendation to the target user. The second is this model can be really very specific to a user by the fact that the recommended items can vary considerably from one user to another. The disadvantage with CBF, is that as the recommendation is made for a user who already has existing interests. Therefore, the ability to expand a user's interests is a limitation which leads to cold-start problem and poor recommendation quality.

2.1.7 Context-Aware Recommender Systems (CARS)

CF is known traditionally as the best approach when it comes to RS. A large number of Web sites and application uses CF because of its simplicity and quality compared to other techniques. However CF has shortcomings related to popularity bias, CF diversity performance is nearly zero [dSPF19]. New items can not be recommended in CF because of the lack of information regarding users (e.g. ratings, implicit feedback). Context-aware recommender systems (CARS) step in to alleviate the cold start problem. In CARS, context is used as dimensions (time, mood, location, etc) and their attributes such as sentiments, country, city, etc which can be used to recommend as it was pointed out by several authors [AT15, PTG08]. The context needs a pre-processing before being used, only the relevant contextual information is used, the process is called contextual pre-filtering. Contextual information is usually categorized in three types namely: explicit when users give the information directly, implicit when users are not aware of the system collecting contextual information, and inferring which uses statistical or other methods in data mining to obtain the contextual information from implicit feedback.

2.1.8 Popularity-Based Recommender Systems (PBRS)

As the name suggests, Popularity-based recommender systems (PBRS) suggest items that are in trend at the moment or overall popular. The system identifies the items that

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

are bought or liked most as popular items or products. For example, if a new user signs up, the system will most likely suggest to her the item that many users are interested in right now. The drawback of popularity based recommender system is that there is no personalization. Even though user behaviour is known, item recommendations are not personalized.

There are other recommender systems we did not mention in detail in this dissertation such as Utility-based recommendation system, Demographic-based recommendation system, Knowledge-based recommendation system, Cross-domain recommendation system. Social-based recommender systems are discussed in Section 2.3.

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

2.2 Social Network Analysis

Figure 2.1: Network example (FilmTrust)

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

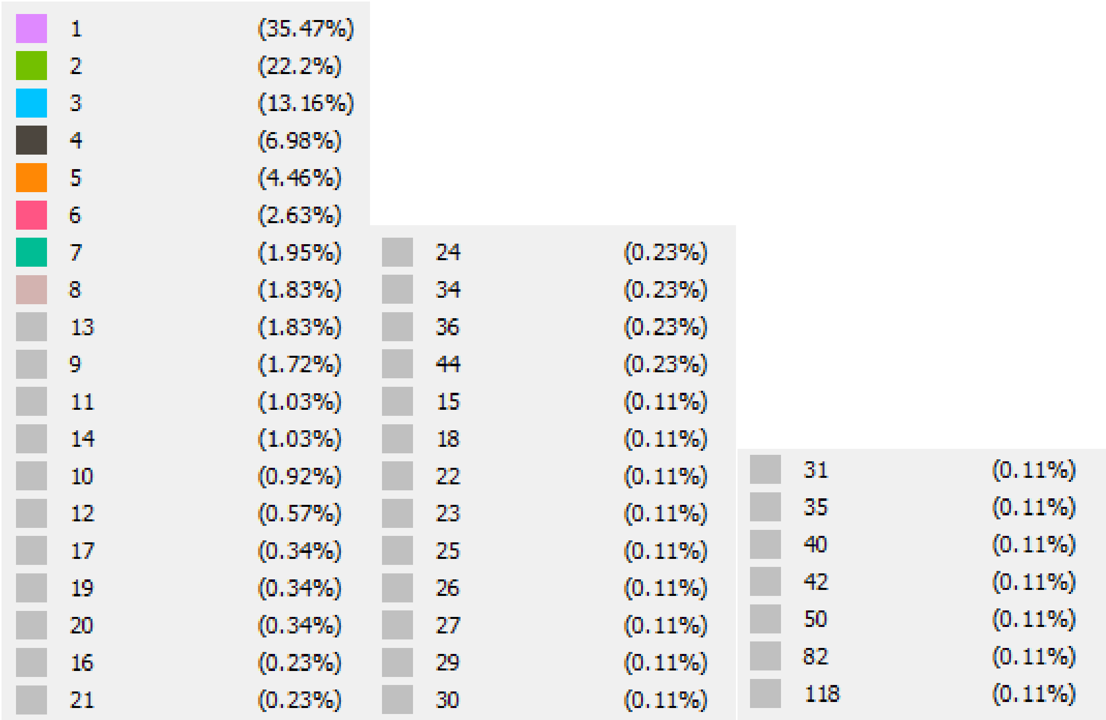


Figure 2.2: Color palettes based on number of connections

Social Networks are sets of nodes connected by edges. In this thesis, each node is considered as a user and each edge as a tie (link). We denote the social network as the graph $G = (U, E)$ where U is the set of users and E is the set of links between them. In most cases while making important decisions or choices users may also turn to their Social Network (SN) for help. In social networks, we have connections (ties) in different ways: with people we share same interests (trust relationships), those we are friends with (social Friendships) and with those we are just connected unwillingly (backgrounds) or unknowingly (activities). Social network analysis (SNA) [Sco00], is the application of the broader field of network science to study human relationships and connections as well as to understand social, political and economic phenomena. The network analysis is mainly done on three different levels, the microscopic approach, that focuses on some nodes of interest, which we tackled at individual level (L1) by degree centrality and PageRank in order to identify and distinguish selected nodes from the rest of the network. At the mesoscopic level, which is the level where the multiplicity of nodes is derived by statistics and distributions are also observed correspond to our second level is covered as well at individual and pairwise level (L2) in this research where the network distance and node similarity is taken into account. The last level macroscopic is also presented in our community level (L3) where community detection approaches are utilized to extract community within the network. Besides the above mentioned and described measures, there exist other measures such as roles distribution [QSZL19] and

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

the network community profile [LLDM08]. FilmTrust is shown as a network example in Figure 2.1 and degree distribution is presented in Figure 2.2 where the percentage shows how many users have a certain amount of connections. For example, 35.74% of users have degree of 1 and 0.11% of users have degree of 118.

2.2.1 Adjancency Matrix (S)

If two nodes are joined by an edge, they are adjacent and we call them neighbors. The adjacency matrix S of an undirected graph with $m = |U|$ nodes/vertices has shape $m \times m$; for an undirected graph, S is symmetric ($S = S^T$). Matrix with elements S_{ik} such that:

$$S_{ik} = \begin{cases} 1, & \text{if there is an edge between } u_i \text{ and } u_k \\ 0, & \text{otherwise} \end{cases}$$

Directed graph:

$$S_{ik} = \begin{cases} 1, & \text{if there is an edge from } u_k \text{ to } u_i \\ 0, & \text{otherwise} \end{cases}$$

2.2.2 Edge and Node Degree

Node degree depends mostly on the type of the network one is dealing with. There are two types of network/graph. Namely directed and undirected graph. In a directed graph, the edges of the network represent a specific direction from one node to another. For example, Phone calls and Twitter. In an undirected network, the edges simply connect a node to the other, it can be via mutual agreement like on Facebook. Degree of a node is the total number of link that are connected to that node.

In-degree and Out-degree

In a directed graph we have two different types of links respectively, in-links and out-links. In-links are the connection pointing to the node and out-links are links from the node to neighbour nodes[TKBK17]. Each node has two degrees: The in-degree d_i^{in} the number of connections coming to a node i; the out-degree d_i^{out} is the number of outgoing edges. The most useful part of in-links and out-links is defining the importance or the popularity of the node. By using the local structure around nodes only degree is considered as the simplest of the node centrality measures. Edge can determine the importance of the node if it is a weighted edge for example in emails.

2.2.3 Centrality in Networks

Each person has some degree of influence or importance within the social domain under consideration, and one expects such importance to surface in the structure of the social

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

network; Centrality is a quantitative method designed to reveal the importance of a node [dABR+14]. In a network some nodes or links are more central or important than others; centrality is a fundamental tool in the study of social networks. Important or prominent nodes are those that are largely linked or related with other nodes (user). It was noted by Freeman [Fre78], that the inferred starting point of all centrality measures is the same: the central node of a popular actor should be more important than the other nodes; ironically, it is precisely the unanimous agreement on this requirement that may have produced quite different approaches to the problem.

Here we consider two definitions of centrality.

Degree centrality

Degree is a simple centrality measure that counts how many neighbors a node has. A vertex is more important when the number of neighbours grows. If the network is directed, we have two versions of the measure: in-degree which is the number of in-coming links, or the number of predecessor nodes; out-degree which is the number of out-going links, or the number of successor nodes. Typically, we are interested in in-degree, since in-links are given by other nodes in the network, while out-links are determined by the node itself. Degree centrality suggests:

A node is important if it has many neighbors, or, in the directed case, if there are many other nodes that link to it, or if it links to many other nodes[1].

The degree centrality of a vertex $u_i \in U$ for a given graph $G = (U, E)$, in terms of the adjacency matrix S is:

$$d_i = \sum_{k=1}^{m} (S_{ki} + S_{ik}).$$

Regarding the social network, degree can give us more information about types of users in the network; users who are "Talkers", where users' out-degree is greater than in-degree number, users who are "Listeners", where users' in-degree is greater than out-degree number. The last group of users are "Communicators", having the same number of in-degree and out-degree.

Pagerank

PageRank is essential when it comes to determining the importance of nodes in the network. It returns a value that indicates the level of importance of a node compared to other nodes by taking into account the importance and relevancy of the connected nodes to the target node. One example of the real noticeable scenario of PageRank is the way a search engine like Google uses PageRank as its trademark to decide which results to display at the top of its search engine listings. The algorithm has been used to determine

[1] https://www.sci.unich.it/ francesc/teaching/network/degree.html

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

the influence level as well in many research as in [LNPC15], they measure the influence of a single article by using its local co-citation network.

PageRank [PBMW99] depends on the number of incoming connections of a user as well as their quality, with higher centrality users giving more importance to their outgoing connections; in some sense, the higher its PageRank is the more respected a user is. In terms of the adjacency matrix S, PageRank satisfies the equation

$$x_i = \alpha \sum_{k=1}^{m} \frac{S_{ki}}{\max\{d_i^{out}, 1\}} x_k + \frac{1-\alpha}{m},$$

where $d_i^{out} = \sum_{k=1}^{m} S_{ki}$ is the out-Degree centrality of user i, and α is the damping factor, typically set to 0.85.

2.2.4 Network Distance and Similarity Measure

Network distance

Network distance is a minimum number of connections or links required to connect two particular nodes, in our case two particular users in the network. It can also be defined as the length of the shortest path between two nodes. In our research, we have used one of the efficient algorithm to compute all pairs' shortest path which is called Floyd-Warshall algorithm. It is used to find the shortest (longest) paths among all pairs of vertices in a graph, it is used for the graph with no cycles of negative length[PM13]. The simplicity of algorithm is one its advantage. Floyd-Warshall algorithm uses a matrix of lengths D^0 as its input. If there is an edge between nodes i and j , then the matrix D^0 contains its length at the corresponding coordinates. The diagonal of the matrix contains only zeros. If there is no edge between edges i and j , then the position $(i,\ j)$ contains positive infinity. In other words, the matrix represents lengths of all paths between nodes that does not contain any intermediate node.

In each iteration of Floyd-Warshall algorithm, this matrix is recalculated, so it contains lengths of paths among all pairs of nodes using gradually enlarging set of intermediate nodes. The matrix D^1 , is the first created matrix during procedure, the matrix contains paths among all nodes using exactly one intermediate node which is determined in advance. D^2 includes lengths using two predefined intermediate nodes. Finally the matrix D^m uses m intermediate nodes. The following recurrent formula describes the above transformation:

$$D_{ij}^{m} = min(D_{ij}^{m-1}, D_{ik}^{m-1} + D_{kj}^{m-1})$$

Network nodes similarity measures: SimRank

The idea behind *SimRank* is simple: two users are similar if they are referenced by similar users [JW02, AGMcC07]. Each user is considered to be completely similar to herself,

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

which gives it a similarity score of 1. The similarity $SR(u, v)$ between users u and v takes values in $[0, 1]$, and satisfies a recursive equation. If $u = v$ then $SR(u, v)$ is defined to be 1. Otherwise,

$$SR(u,v) = \frac{C}{|N(u)||N(v)|} \sum_{u' \in N(u)} \sum_{v' \in N(v)} SR(u', v'),$$

where C is a constant between 0 and 1, and u', v' are in-neighbors of users u and v, belonging to the sets $N(u)$ and $N(v)$, respectively. A detail here is that either u or v may not have any in-neighbors. Since there is no way to assume any similarity between u and v in this case, SimRank is set to $SR(u, v) = 0$, which makes the addition of the main equation to be 0 when $N(u) = \emptyset$ or $N(v) = \emptyset$. SimRank can be considered as a global pairwise similarity measure.

Network nodes similarity measures: LHN (Leicht Holme Newman)

Leicht Holme Newman index [GZ16, LHN06] counts the expected number of common neighbors between two users. For users u and v the LHN is computed as:

$$LHN(u,v) = \frac{|N(u) \cap N(v)|}{d_u \times d_v},$$

where $N(u)$ is the neighborhood of user u, and d_u is the degree of u. Intuitively, LHN assigns a high similarity score to pairs of users that have many common neighbors [Sch15]. LHN, in contrast to SimRank, can be considered as a local pairwise similarity measure.

2.2.5 Community Detection Approaches

Community detection is a complex issue in the network analysis theory for the reason that the networks are also complex. Communities are very important when it comes to understanding the function and structure of social networks. Optimization of the quality function modularity known as modularity-based is the method often borrowed from other community detection, many researchers such as [New06, New16] have proven the effectiveness of this algorithm in several complex networks. A community is characterized by users who have a common point; to understand the grouped large number of users very well we decided to use community detection approaches and analyze their connectivity behaviour within the communities. Two different methods to extract communities are considered namely: Influencer based communities (inf-communities) and Modularity based communities (mod communities). The following methods are elaborated below

Influencer based communities (inf-communities)

This method is based on the level of influence which we determined by degree centrality, users who have the degree centrality value higher than a threshold are categorized as influencers. Each central user is chosen based on degree centrality, the community is formed by all users that are connected to the influencer.

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

Modularity based communities (mod-communities)

This method is based on greedy modularity community detection algorithm [New06]. In the first step is each node is considered as a community or its own cluster. In the second two clusters are merged which will increase the modularity by the highest number. We stop once step two (all merges) would reduce the modularity. The modularity is used as a measure the quality of clustering. Modularity provides information about how the communities are formed within a network. It measures the strength of the division of a network into communities/groups. High modularity indicates that the network has dense connections between nodes within communities but sparse connections between nodes in different groups.

2.3 Social Recommender Systems

Traditionally, people turn to their friends, or those whose opinions they trust, when looking for advice and recommendations on a specific domain. *Social Recommender Systems* attempt to mimic this behavior by also drawing information from the social context of the users. The underlying assumption is that the decision process of a user depends not only on her individual preferences, but also on influence she receives from her social connections. This is motivated by the phenomena of homophily and influence in social networks [EK10]; the former suggests that users socially connect because they have similar interests, while the latter says that socially connected users tend to develop similar interests. Specifically, social recommenders exploit two distinct sources of information, the historical rating behavior of users, exactly like collaborative filtering techniques, and the social network, and assume that these sources are correlated, implying that the latter can provide additional information about the former. As an example, consider the role of the neighborhood in user-user collaborative filtering, e.g., in [RIS+94]. The users in the *rating neighborhood* of a target user essentially provide additional knowledge about the preferences of the target user, so that the system can make better informed recommendations. In analogy, in social recommenders, e.g., [MA07], the users connected to the target user, i.e., her *social neighborhood*, also provide complimentary information about the preferences of the target user. Social recommendations is an active research area in the past few years, e.g., refer to [MA07, MYLK08, MKL09, BAX10, MZL+11, ABS+12, LA13, LWTM15, Guy15, AV16]. Social recommender systems borrow ideas from *Collaborative Filtering* (CF), which is the most commonly used method for making recommendations. In CF approaches, users and items with similar rating patterns are taken into account [JCS04] to produce a recommendation for the target user. The meaning of a recommendation varies depending on the concrete social network platform. In platforms such as Facebook a link between two people is established if both persons agree to have a friendship relation [Sch15]. The resulting network is thus un-directed because both persons share mutual friendship. Another example of a social network is Twitter. In Twitter a link between two persons is established if a user is interested in news updates of another user.

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

2.3.1 Trust aware Recommender systems (TaRS)

Trust is very useful in our daily life when it comes to choosing friends and to what extent to trust them. Trust aware recommender system is based on traditional CF, the concept of trust propagation is taken into account; the system predicts the trust value between two users even if it has not been explicitly stated. Trust plays a big role while predicting the ratings. Trust can be explicit or implicit, when it is given directly by a user to another user it is explicitly expressed and in that case, it is explicit trust, it can be in form of scale for example from 0 to 1. Implicit trust is defined by user behavior on the system; it can be determined by using similarity metrics. Recently researchers have been working on improving TaRS such as [NGT12] by using data including user ratings on different items and also trust/distrust of users on each other. Trust network is constructed based on trust/distrust information by using the similarity of users which is calculated from the Pearson correlation coefficient with fixed thresholds. They increased accuracy by dropping the trust edge between two users where similarity falls below the threshold. The idea of Trust-Aware recommender system (TARS) is very similar to CF recommender systems.

The user based rating prediction formula:

$$p(u_i, i_j) = \bar{r}_{u_i} + \frac{\sum_{u_f \in N_{u_i}} W_{u_i,u_f} \times (r_{u_f,i_j} - \bar{r}_{u_f})}{\sum_{u_f \in N_{u_i}} W_{u_i,u_f}}$$

where N_{u_i} represents the set of top-k similar users; $\bar{r}_{u_i}$ the average ratings of user u_i; r_{u_f,i_j} the rating of user u_f to item i_j; W_{u_i,u_f} is the similarity weight of users u_i and u_f.

Trust aware Recommender systems (TaRS)[MA07] The model utilizes trust neighborhoods in different distances from the target user, Trust Neighborhood (TN1) is when the estimated trust value is predicted from the distance 0 (the target user) to the user's direct friends (distance 1), TN2 where users (friends of friends) are taken into account to estimate the trust value in order to predict rating of the target user u_i. Let TN_{u_i} be a set of users in trust network. u_i the target user.

$$TN_{u_i} = \{TN_{u_{i_1}} \subset TN_{u_{i_2}} \ldots \subset TN_{u_{i_k}}\}$$

Where $1 \ldots k$ represents the distance from user u_i to some user in the network. To predict rating, formulas depend most on the distance. Below is the prediction formula, considering users who are direct friends to user u_i:

$$p(u_i, i_j)_{TN_{u_{i_1}}} = \bar{r}_{u_i} + \frac{\sum_{u_f \in TN_{u_{i_1}}} T_{u_i,u_f} \times (r_{u_f,i_j} - \bar{r}_{u_f})}{\sum_{u_f \in TN_{u_{i_1}}} T_{u_i,u_f}}$$

Where $TN_{u_{i_1}}$ represents the prediction of user u_i to item i_j by taking user u_f direct friend of u_i in trust network. T_{u_i,u_f} is the estimated trust value between u_i and u_f. The

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

prediction by considering friends of friends is:

$$p(u_i, i_j)_{TN_{u_{i_2}}} = \bar{r}_{u_i} + \frac{\sum_{u_g \in TN_{u_{i_2}}} T_{u_i,u_g} \times (r_{u_g,i_j} - \bar{r}_{u_g})}{\sum_{u_g \in TN_{u_{i_1}}} T_{u_i,u_g}}$$

The conclusion based on results is that the best accuracy is produced by TN1, because as the distance increases the model performance becomes poor. The coverage which determines how much the model is able to produce predicted rating increases with the distance from the target user. Trying to combine the traditional CF with their model did not give good results.

2.3.2 Social Recommandations (SoRec)

SoRec [MYLK08] is one of extended basic MF model that incorporate the social network. S is the matrix representation of the social network; $I_{if}^S = 1$ if users u_i and u_f are friends and $I_{if}^S = 0$ otherwise. Matrix S is factorized into a user-specific matrix U and a factor-specific matrix F. The latent feature vectors of users are learnt based on both the rating and social network matrices. The objective function to be minimized is:

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}(R_{ij} - U_i^T V_j)^2 + \frac{\lambda_S}{2}\sum_{i=1}^{m}\sum_{f=1}^{m} I_{if}^S (S_{if} - U_i^T F_f)^2$$

$$+\frac{\lambda_U}{2}\|U\|_F^2 + \frac{\lambda_V}{2}\|V\|_F^2 + \frac{\lambda_F}{2}\|F\|_F^2$$

Where $\lambda_F, \lambda_V, \lambda_U, \lambda_S$ are regularization weights and $\|.\|_F^2$ denotes the Frobenius norm. In SoRec λ_S is a very important parameter which balances the user-item rating matrix and the user social network. With a certain threshold the increase of λ_S the prediction accuracy decrease. The larger λ_S is the better the model performs.

2.3.3 Social Trust Ensemble (STE)

Social trust ensemble [MKL09] model is a linear combination of basic MF and a social network based approach (Trust based recommendation technique). STE modified the basic MF model so that each rating R_{ij} is the aggregate of the estimated ratings of friends/neighbors. The objective function is:

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}\Big(R_{ij} - (\alpha U_i^T V_j + (1-\alpha)\sum_{u_f \in N(i)} S_{if} U_f^T V_j)\Big)$$

$$+\frac{\lambda_U}{2}\|U\|_F^2 + \frac{\lambda_V}{2}\|V\|_F^2$$

Where S is the matrix representation of the social network, $N(i)$ is the friend set of user u_i and α controls the effect of friends on the rating estimation. The equation $U_f^T V_j$ is

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

the estimate rating of neighbour u_f on j. S_{if} is the normalized trust value between u_i and u_f. In binary trust networks, it would be one over the number of friends. When it comes to deal with cold start users STE is limited because those users ratings are very few [JE10]. And to be able to learn the features using STE model enough ratings should have been expressed by the user.

2.3.4 RS with Social Regularization (RSR)

RSR [MZL+11] is defined as using social friends network to improve recommender systems, a way to model social network information as regularization terms to constrain the matrix factorization framework. In this model the training can be done on two types of datasets: dataset from social friend network and dataset from trust network (web of trust).

Average based Regularization In the real world it is believed that our friends recommendations will have a big impact on our choices and decisions because we believe in the tastes and suggestions of our friends. Based on this intuition, the first social recommendation model in SR is based on the matrix factorization technique. The objective function to be minimized is:

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}(R_{ij} - U_i^T V_j)^2 + \frac{\alpha}{2}\sum_{i=1}^{m} \|U_i - \frac{\sum_{u_f \in N(i)^{in}} Q_{if} \times U_f}{\sum_{u_f \in N(i)^{in}} Q_{if}}\|_F^2 + \frac{\lambda_U}{2}\|U\|_F^2 + \frac{\lambda_V}{2}\|V\|_F^2$$

Where λ_V, λ_U are regularization terms and $\|.\|_F^2$ denotes the Frobenius norm. $N(i)$ is the friend set of user u_i, $N(i)^{in}$ denote u_i's outlink friends, this applies when the network is directed, for example, Filmtrust network. The undirected network like Douban $N(i)^{in} = N(i)^{out}$, therefore $N(i)$ can be used alone. $\alpha > 0$ and Q_{if} is the similarity value between user u_i and user u_f from matrix similarity Q given by adjusted cosine similarity normalized to [0,1].

Individual based Regularization This method is quite similar to the average based regularization, the only difference is that it treats one user and her friends individually, assuming that some of user u_i friends have diverse tastes. For this social recommendation model the objective function to be minimized is:

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}(R_{ij} - U_i^T V_j)^2 + \frac{\beta}{2}\sum_{i=1}^{m}\sum_{u_f \in N(i)^{in}} Q_{if}\|U_i - U_f\|_F^2 + \frac{\lambda_U}{2}\|U\|_F^2 + \frac{\lambda_V}{2}\|V\|_F^2$$

where $\beta > 0$, and Q_{if} is the same similarity as the one used in the previous equation. When Q_{if} value is small, it shows that the distance between feature vectors U_i and U_f should be larger, otherwise the distance between the feature vectors should be smaller.

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

2.3.5 MFC and MFC$^+$

MFC and MFC$^+$ models [LWTM15], insert community interest constraints into the SR model by treating users who belong to different communities differently unlike RSR model, which treats them likewise. If the target user has one community from which he is more interested in than another, the community he is more interested in should be weighed higher. The distance between the target user and his/her interested communities is minimized by MFC$^+$ model, while in the same community the distance between the target user and other members of the community is minimized by MFC [LWTM15]. MFC and MFC$^+$ consider two kinds of cold-start users: rating cold start users (with few ratings) and social cold-start users (with few social ties).

2.4 Stastisical Tools

2.4.1 ANOVA

ANOVA statistical test was used in this study to compare and observe the behavioural changes or impact across different groups that were obtained through partitioning. However, before applying ANOVA there are several assumptions to be satisfied[Saw09]. For example, interval data of the dependent variable, normality, homoscedasticity, etc. Most statistical tests rely on the normality of a sample data; it is indispensable to test whether the basic distribution is normal, or at least symmetric. This test can be conducted by using different approaches; the common ones include the graphical distribution review by histograms, QQ plots and boxplots. These approaches allow detecting whether there are any outliers within the partitions. In this study, two functions in R were used to create Q-Q plots: qqnorm and qqplot. The Homogeneity of Variances test was also done (homoscedasticity)[KC18]. The Levene's test was done for every three partitions/groups used during experiments. The hypothesis used in the test should not be significant (not small p-value) to meet the assumption of equality of variances. If the assumption does not hold, it is advisable to do the ANOVA test with Welch correction.

In our research, we have three groups A, B, C (corresponding to low, medium, high values of some attribute) and the hypotheses of interest in an ANOVA are as follows:

Hypothesis 1: $H0 : \mu_A = \mu_B = \mu_C$

Hypothesis 2: *H1: Means are not all equal.*

ANOVA shows whether results are significant overall, but it does not reveal exactly where those differences lie. Therefore, if the ANOVA test is positive, we perform Post hoc analysis which is more elaborate test, that allows for unequal sample sizes, and lack of variance homogeneity. Tukey's HSD [Tuk49] or Dunett's T3 [Dun80] test is applied on partition pairs e.g. B-A, C-B, C-A, to check whether a trend in the test attribute is significant.

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

TUKEY's Test (Tukey's HSD)

Tukey's Honestly Significant Differences (HSD) compares pairs of the sample means by using their absolute differences. Much of the work on multiple comparisons has been based on the original work of Tukey, and an important test bears his name. The Tukey criterion D_{Tukey} is defined as:

$$D_{Tukey} = q_{\alpha(c,n-c)}\sqrt{\frac{MSE}{n_i}}$$

where $q_{\alpha(c,n-c)}$ represent Studentized range distribution based on c and degree of freedom df, c number of column, n total sample size and n_i sample size of the sample group with the smallest number of observations. MSE is the mean square error within the groups from ANOVA table.

DUNNETT's Test (Dunett's T3)

In some experiments the important comparisons are between one control group and each of several experimental group. In this case, the most appropriate test is Dunnett's test. Dunett's is used to compare a simple group mean against all other group means. It compares the control group to all of experimental group means. It acts like t test but comparing two groups. It is given by the formula:

$$D_{Dunnett} = t_{Dunnett}\sqrt{\frac{2MS_{S/A}}{n}}$$

where $D_{Dunnett}$ stands for the difference, to determine $t_{Dunnett}$ or t_d we need to know how many groups we have, how many degree of freedoms $df_{S/A}$ among groups and what is alpha. We will let t_d represent the critical value of a modified t statistic. The critical value t_d is found in tables supplied by Dunnett (Dunnett-critical value table). $MS_{S/A}$ is the mean squares (MS) of the within group in the ANOVA source table. A standard t test between the appropriate means can also be utilized (using MSE as the variance estimate and evaluating the t against the tables of t_d) or solve for a critical difference between means. For a difference between means μ_A and μ_B (where μ_A represents the mean of the control group) to be significant, the difference must exceed $D_{Dunnett}$.

Pearson and Spearman Correlation

The use of correlation measures in this dissertation is to measure the extent to which two attributes tend to change together from one view to another. The Pearson correlation is a correlation measure that evaluates the linear relationship between two continuous variables while Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables. When evaluating the relationship between two attributes, it is important to determine how the attributes are related. Therefore, two different

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

types of correlation result can be determined from Pearson and Spearman correlations. A positive linear relationship is nearly perfect and also called a strong correlation when both attributes increase concurrently and at a constant rate. In increasing order the relationship is positively correlated while a negative linear relationship exists when one variable increases while the other variable decreases.

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

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

CHAPTER 3

Methodological Approach

We investigate whether there exist relationship/correlations between users based on social ties and based on ratings behaviour at different levels (level of users, level of user to user and level of communities).

This chapter presents the attributes used to describe the objects of study, users at L1, pairwise at L2 and communities at L3. At the first level, the attributes capture the level of *activity* of users in terms of their rating behavior and social connections. At the second level, the attributes capture the *similarity* between two users, again in terms of their rating behavior and social connections. At the third level we have two methods which we used to extract communities from the network.

Section 3.1 presents the attributes used and methods at L1, it discusses also how we studied correlations. Section 3.2 elaborates attributes used, methods at L2. Section 3.3 elaborates attributes used, methods at L3 and Section 3.4 presents the datasets used in our study.

3.1 Methods at Individual Level

3.1.1 Attributes Capturing Activity of Users (L1)

One notion of activity in terms of rating behavior was considered, and two notions in terms of social connections, based on the concept of node centrality [BV14].

RATE-NUM. For the rating behavior, the number of ratings a user has provided. denoted as RATE-NUM was considered. This essentially, captures how "*heavy*" rater a user is.

NET-DEG. Degree is the most intuitive interpretation of popularity, as it counts the number of (incoming or outgoing) connections a user has.

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

NET-PR. PageRank [PBMW99] depends on the number of incoming connections of a user as well as their quality, with higher centrality users giving more importance to their outgoing connections; in some sense, the higher its PageRank is the more respected a user is.

3.1.2 Partitioning Method

Objects were divided into three partitions; A, B and C according to an attribute of one view, called the *partition attribute*. Partition A contains objects with low activity (users in L1), B contains users with medium activity (neither high or low) while partition C contains objects with high activity. In each partition, the mean of an attribute of the other view, called the *test attribute* was computed. Then the test attribute was examined to see whether it increases with the partition attribute. For example, for partitions based on RATE-NUM, the average NET-DEG (or NET-PR) was computed, and then observed whether the mean NET-DEG increases from partition A through C. To formally test if there is a trend, ANOVA test was done to investigate whether the mean of the test attribute is significantly different across partitions. ANOVA test shows whether the results are significant overall, but it does not reveal exactly where those differences lie. Therefore, for a positive ANOVA test result, a post hoc analysis, Tukey's HSD was performed [Tuk49] or Dunett's T3 [Dun80] test, on pairwise differences of the partition means (B-A, C-B, C-A), to check whether a trend in the test attribute is significant.

3.1.3 Ranking Method

For the second approach, termed *ranking*, we create a rankings of objects based on each attribute, and retain only those that have the highest activity (L1); the selected attribute is called the *ranking attribute*. For example, we may construct the ranking of the top-100 most heavy raters (RATE-NUM) in the system. Then, we look for correlations between the two views (ratings and social behavior) in two ways. In the first way, we pick two rankings produced by attributes of different views, and count the number of common objects in them. For example, we see how many users are both among the top-100 most heavy raters (RATE-NUM) and the top-100 most well connected users (NET-DEG). In the second ranking method, we pick one ranking, and study the correlation, measured by Pearson's and Spearman's correlation coefficients, between two attributes of different views. For example, we select the 100 most active users according to their ranking (RATE-NUM), and see how their RATE-NUM correlates to NET-DEG.

3.2 Methods at Pairwise Level

3.2.1 Attributes Capturing Similarity Between Two Users (L2)

We consider one notion of similarity in terms of rating behavior, and three notions in terms of social connections.

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

RATE-SIM. The pairwise cosine similarity metric finds the normalized dot product of the rating vectors of two users [SKKR01]. This simple definition however, has some limitations. It is known that people tend to rate on different scales. Some people are naturally high raters, which means they might rate items highly in general, even if they do not like the item very much. There are some people who tend to rate low, even when they like the items very much. The traditional cosine similarity does not consider the difference in rating scale among different users [LH16]. The adjusted cosine similarity offsets this drawback by subtracting the corresponding user average from each co-rated pair. Formally, the similarity, denoted as RATE-SIM, that was used between users u and v is given by:

$$sim(u,v) = \frac{\sum_{i \in I_u \cap I_v}(r_{ui} - \bar{r}_u).(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_u}(r_{ui} - \bar{r}_u)^2}.\sqrt{\sum_{i \in I_v}(r_{vi} - \bar{r}_v)^2}},$$

where I_u and I_v are the sets of items rated by user u and v, r_{ui} is the rating user u gave to item i and $\bar{r}_u$ the average of all ratings given by u.

RATE-PCC. Pairwise similarity (RATE-PCC) is the rating similarity when only the common rated items between two users are considered:

$$sim(u,v) = \frac{\sum_{i \in I_u \cap I_v}(r_{ui} - \bar{r}_u).(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_u \cap I_v}(r_{ui} - \bar{r}_u)^2}.\sqrt{\sum_{i \in I_u \cap I_v}(r_{vi} - \bar{r}_v)^2}},$$

where I_u and I_v are the sets of items rated by user u and v, r_{ui} is the rating user u gave to item i and $\bar{r}_u$ the average of all ratings given by u. The $\sum_{i \in I_u \cap I_v}$ is the sum over items that both users have rated in common.

The difference between RATE-SIM and RATE-PCC is that RATE-SIM sums the items that users rated individually while RATE-PCC sums the items that both users have rated in common.

RATE-JACC. The Jaccard similarity index (also called the Jaccard similarity coefficient or the Tanimoto index/coefficient) is a popular measure for similarity between two sets of binary data. In our case, we measure similarity between two users by considering the sets of items they have interacted with (implicit feedback). RATE-JACC computes the cardinality ratio of the intersection and the union of the rated (or interacted with) items:

$$J(u,v) = \frac{|I_u \cap I_v|}{|I_u \cup I_v|}.$$

where I_u and I_v are the sets of items rated by user u and v.

NET-DIST. Network distance between two users is a minimum number of connections, or links, that separate them in the network. It can also be defined as the length of the shortest path between two users. The algorithm of Floyd-Warshall[PM13] was used to determine the network distance of all pairs of users.

NET-SIM. NET-SIM is the SimRank discussed in Section 2.2.

NET-LHN. NET-LHN is the *Leicht Holme Newman* index discussed in Section 2.2.

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

3.2.2 Partitioning and Ranking Methods

The partitioning method was also applied at pairwise level (L2); objects were divided into three partitions, A, B and C according to an attribute of one view, called the *partition attribute*. Partition A contains objects with low similarity (pairs of users in L2), B contains users with medium similarity (neither high or low) while partition C contains objects with high similarity. In each partition, the mean of an attribute of the other view, called the *test attribute* was computed. Then the test attribute was examined to see whether it increases with the partition attribute. For example, for partitions based on RATE-SIM, the average NET-SIM (or NET-LHN) was computed, and then observed whether the mean NET-SIM increases from partition A through C.

In the ranking method, we pick one ranking, and study the correlation, measured by Pearson's and Spearman's correlation coefficients, between two attributes of different views. For example, we select the 100 and 1000 most similar pairs according to their ranking (RATE-SIM), and see how their RATE-SIM correlates to NET-SIM.

3.3 Methods at Community Level

3.3.1 Attributes Capturing Similarity Between Two Users (L3)

Matrix Factorization Similarity based method (MF-SIM) Matrix factorization similarity (MF-SIM) method was used. The prediction is given by the dot product of UL and IL. MF-SIM computes the similarity for every user to all others by computing inner products of user's lantent factors.

3.3.2 Influencer based communities

Influencer based communities mechanism focuses more on users with a degree centrality greater than 10, we use term influencers to distinguish these users with the rest of other users in the network. Each community is formed by friends of an influencer. *inf_community* operates in a way the central user satisfies the condition of influencers and all other users connected to the central users form a community, within one inf_community it is possible to find many influencers.

3.3.3 Modularity based communities

The optimization of *"Modularity"* which is the quality function is one of the most effective mechanisms for possible partition in the social network and networks in general. The important key is modularity when it comes to mod_communities. *"Modularity"* identifies groups embedded within a network by locating sets of nodes that interact with each other more frequently than the rest.

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

3.4 Dataset

Table 3.1: Datasets basic description

Data Set	Description				Links			Contexts
	Users	Items	Ratings (scale)		Users	Link (Type)		Items
Ciao DVD	7,375	99,746	278,483	[1, 5]	7,375	111,781	Trust	General
FilmTrust	1,508	2,071	35,497	[0.5, 4.0]	1,642	1,853	Trust	Movie
Douban	129,490	58,541	16,830,839	[1, 5]	129,490	1,692,952	Friendship	Movie

In this thesis, three publicly available datasets collected from traces of user interaction in social recommenders were used as shown in Table 3.1. These datasets are commonly used in different studies that have been reviewed in the literature and contain rating activity, i.e., a ratings matrix R, as well as information about the social connections among users, i.e., an adjacency matrix S.

The first dataset, FilmTrust[GZY13], comes from a social networking site in which users can rate and review movies.[1] FilmTrust essentially contains two sub-datasets, a social network in addition to the user-item ratings. The social connections are bidirectional and capture the trust between users (trustee, trustor). Users can specify a level of trust from 1 to 10. However, due to its sharing policy, the exact level of trust is not available, and we only know whether such a connection exists or not.

FilmTrust contains 1,508 users, 2,071 items, 35,497 ratings, and 1,853 social connections. As there exist 635 users with no social connections, and 133 with no ratings history, these were excluded from the analysis. Therefore, only 740 users that have rated at least one item and trust, or are trusted by, at least another person were used in this study. The mean number of ratings per user was 23.5 with the minimum and the maximum being 1 and 244 respectively. The ratings scale is from 0.5 to 4 with a step 0.5, and the mean rating score over all ratings is 3.0.

The second dataset CiaoDVD [GZTY14], is collected on the Ciao website.[2] CiaoDVD contains the social connections among its users. Compared to FilmTrust, CiaoDVD is about an order of magnitude larger, having 17,588 users, 16,121 items, 72,665 ratings, and 40,133 social connections. However, there exist 12,930 users with no social connections, and 1,918 with no ratings history. These were excluded from the analysis done in this study, and only 2,620 users with both pieces of information. The mean number of ratings per user is 12.57 with the minimum and the maximum being 1 and 1,106 respectively. The ratings scale is from 1 to 5, and the mean rating score over all ratings is 4.07.

[1] http://trust.mindswap.org/FilmTrust
[2] http://dvd.ciao.co.uk

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

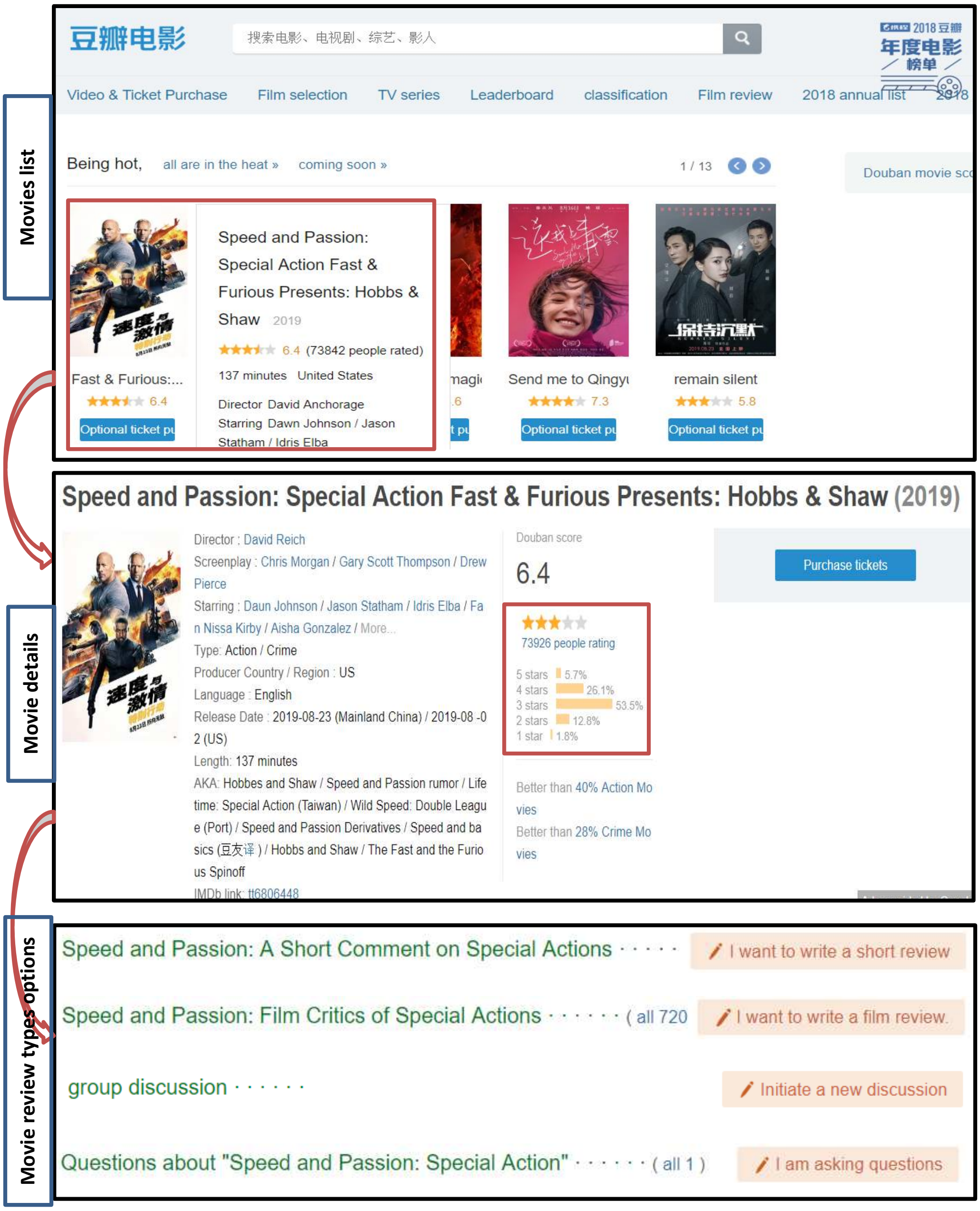


Figure 3.1: Example: Douban online movies

The third dataset DOUBAN (Figure 3.1)[3]. Douban dataset [RSK16, MB15]. Douban

[3]https://movie.douban.com/

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

contains 129,490 unique users and 58,541 unique movie items. The total number of movie ratings is 16,830,839. For the social friend network, there are a total of 1,692,952 social relationships. Douban Movie is a Chinese website that allows Internet users to share their comments and viewpoints about movies. Users are able to post short or long comments on movies and give them marks. Two files are included in this Douban dataset, the user-item rating file (UserId, ItemId, Rating) and the user social friend network file(UserId1, UserId2). When this dataset was crawled, Douban only allowed the Facebook-like friendship building mechanism (Now Douban also supports the Twitter-like following mechanism).

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

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

CHAPTER 4

Analysis at Individual Level

Social recommender systems make use of the available information about social connections between users to improve the quality of the recommendations. The assumption is that if two users are connected, they are likely to have similar preferences, and thus the system should make similar recommendations. Recently many approaches have been proposed based on similar assumptions, whose validity however has not been systematically studied. In this study, we made the first step towards examining whether there exists observable relationships between social connections and rating behavior in social recommenders. In particular, we examine publicly available datasets (FilmTrust and Ciao DVD) containing traces of rating behavior along with a social graph. We address our first research question:

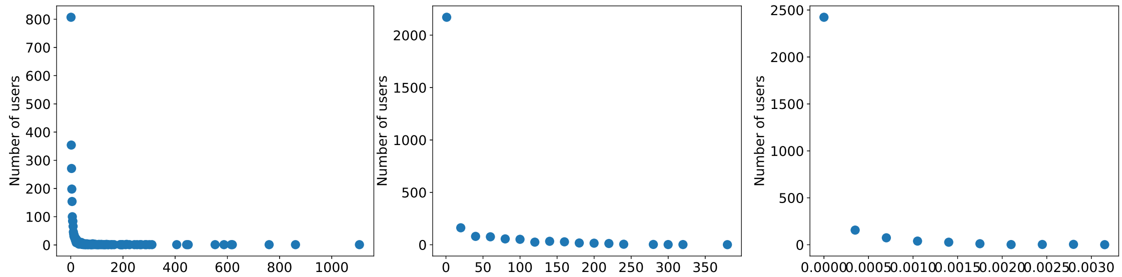
RQ1 Does high activity of the user in one view imply high activity in the other view?

Using techniques from social network analysis and statistics, we investigate whether heavy rates, having provided feedback on many items, are also popular, i.e., central in the social network, and vice versa. We answer several sub-questions such as *Are heavy raters popular?* Our results indicate important connections between heaviness and popularity. Specifically, we find that heaviness implies popularity, and that the association is stronger among very heavy raters.

4.1 Attributes

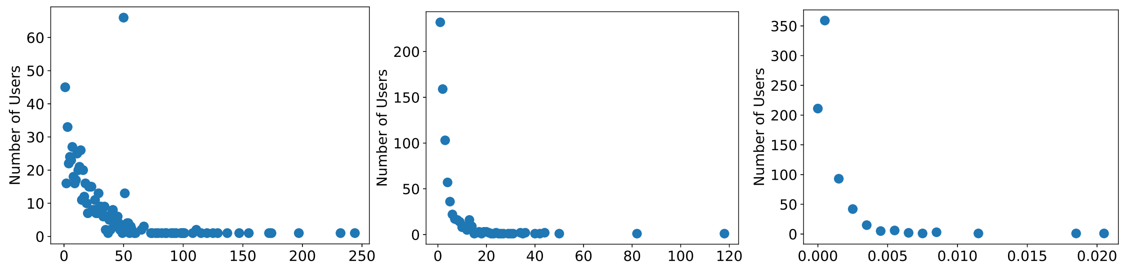
We define "*heaviness*" of a user in terms of the number of ratings (RATE-NUM) s/he has provided. Moreover, we define "*popularity*" of a user as the *centrality* of the node representing the user in the social graph. Every person has some degree of influence or importance within the social domain under consideration, and one expects such

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



(a) RATE-NUM (b) NET-DEG (c) NET-PR

Figure 4.1: Probability distribution of a user having specific values of RATE-NUM, NET-DEG, and NET-PR (CiaoDVD)



(a) RATE-NUM (b) NET-DEG (c) NET-PR

Figure 4.2: Probability distribution of a user having specific values of RATE-NUM, NET-DEG, and NET-PR (FilmTrust)

importance to surface in the structure of the social network; centrality is a quantitative measure that aims at revealing the importance of a node [BV14]. Two definitions of centrality are considered in this study.

Figure 4.2 shows three probability distributions in Filmtrust. First, Figure 4.2a depicts the probability (in raw numbers) of a user being heavy, i.e., giving a specific number of ratings, which we hereafter refer to as RATE-NUM. Then, Figure 4.2b shows the probability of a user being popular in terms of Degree; the mean Degree is 4.7, with min and max values of 1 and 118. Figure 4.2c draws the probability of a user being popular in terms of NET-PR; the mean NET-PR is 0.0012, with the min and max values of 0 and 0.21. These right-skewed distributions show that the majority of users give few ratings and have low centralities, and that at the same time there exist several users that are very heavy and very popular. Figure 4.1 presents the same distributions for CiaoDVD. The mean NET-DEG is 21.75, with min and max values of 1 and 349, while the mean NET-PR is 0.000241, with min and max values 0, 0.003440.

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

Table 4.1: Description of Partitions (FilmTrust)

	A	div	B	div	C
RATE-NUM	242	11	267	30	231
NET-DEG	232	1	262	4	246
NET-PR	241	5.4×10^{-4}	268	1.1×10^{-3}	231

Table 4.2: Description of Partitions (ciaoDVD)

	A	div	B	div	C
RATE-NUM	808	2	978	6	957
NET-DEG	899	2	930	6	914
NET-PR	913	1.3×10^{-4}	913	1.6×10^{-4}	914

4.2 Partition-Based Analysis

To assess the relationship between RATE-NUM and centralities, we consider three distinct divisions, one per each attribute, RATE-NUM, NET-DEG, NET-PR. A division splits users into three partitions, A, B, C, in increasing value of the partitioning attribute. We first determine the lower and upper terciles (3-quantiles) of the partitioning attribute and divide accordingly. Partitions are thus balanced, with each containing roughly 1/3 of all users. Descriptions of the partitions are shown in Tables 4.1 and 4.2 for FilmTrust and CiaoDVD, respectively.

Does the mean NET-DEG differ across RATE-NUM partitions? In the first experiment, we partition users according to RATE-NUM, and compute the mean NET-DEG in each partition. Then, we apply ANOVA to investigate whether the mean NET-DEG is significantly different across partitions. The results for FilmTrust is shown in the top part Table 4.3, where an F value of 24.4 provides significant evidence against the hypothesis that the means are equal (p-value in the order of 10^{-11}).

Following this result, we investigated whether the mean NET-DEG increases from partitions A through C. The Tukey HSD test was applied to check the difference in the mean degree of every pair of partitions is significant. The difference of means and its corresponding 95% confidence interval for each pair are shown in the bottom part of Table 4.3. As suspected partitions A and B, containing non-heavy users, have mostly similar mean NET-DEG and no significant difference is observed. However, there is a significant difference when we compare either A or B with partition C of heavy raters.

Results of ANOVA and Tukey HSD test for CiaoDVD are shown in Table 4.4, where similar conclusions can be drawn. In general, heavy raters tend to be more popular (in terms of Degree) compared to others.

Does the mean RATE-NUM differ across NET-DEG partitions? We also study the reciprocal association. The ANOVA analysis based on the mean RATE-NUM among partitions based on NET-DEG is shown in Table 4.5, where an F value of 11.49 provides

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

Table 4.3: ANOVA and Tukey Test on Mean NET-DEG among RATE-NUM Partitions (FilmTrust)

	DF	Sum. Sq.	Mean Sq.	F value	Pr (>F)
partition	2	2942	1470.9	24.4	$\mathbf{5.18 \times 10^{-11}}$
Residuals	798	48102	60.3		

Pair	Diff. of Means	95% CI	p-value
B - A	0.861	$[0.716, 2.439]$	0.40
C - B	3.565	$[1.987, 5.143]$	$\mathbf{4 \times 10^{-6}}$
C - A	4.427	$[2.849, 6.004]$	$\approx \mathbf{0}$

Table 4.4: ANOVA and Tukey Test on Mean NET-DEG among RATE-NUM Partitions (ciaoDVD)

	DF	Sum. Sq.	Mean Sq.	F value	Pr (>F)
partition	2	1969	984.6	340.3	$\mathbf{2 \times 10^{-16}}$
Residuals	2928	8472	2.9		

Pair	Diff. of Means	95% CI	p-value
B - A	1.08	$[-3.75, 5.92]$	0.86
C - B	19.80	$[14.96, 24.65]$	$\approx \mathbf{0}$
C - A	20.89	$[16.05, 25.73]$	$\approx \mathbf{0}$

Table 4.5: ANOVA and Tukey Test on Mean RATE-NUM among NET-DEG Partitions (FilmTrust)

	DF	Sum. Sq.	Mean Sq.	F value	Pr (>F)
partition	2	16018	8009	11.49	$\mathbf{1.21 \times 10^{-5}}$
Residuals	783	545748	697		

Pair	Diff. of Means	95% CI	p-value
B - A	-0.313	$[-5.729, 5.103]$	0.99
C - B	9.729	$[4.317, 15.145]$	$\mathbf{8.1 \times 10^{-5}}$
C - A	9.416	$[3.91, 14.832]$	$\mathbf{1.4 \times 10^{-4}}$

significant evidence against the hypothesis that the means are equal (p-value in the order of 10^{-5}. The Tukey HSD test shows that partitions A and B of non-popular users have mostly similar mean RATE-NUM and no significant difference is observed. However, there is a significant difference between B and C, and between A and C, implying that popular (in terms of Degree) users tend to be heavier raters. Results on CiaoDVD, Table 4.6, suggest an identical relationship.

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

Table 4.6: ANOVA and Tukey Test on Mean RATE-NUM among NET-DEG Partitions (ciaoDVD)

	DF	Sum. Sq.	Mean Sq.	F value	Pr (>F)
partition	2	248535	124268	55.21	$\mathbf{2 \times 10^{-16}}$
Residuals	2784	6265827	2251		

Pair	Diff. of Means	95% CI	p-value
A - B	2.54	$[-2.43, 7.52]$	0.45
C - B	17.24	$[17.24, 12.26]$	$\approx \mathbf{0}$
C - A	19.79	$[14.81, 24.77]$	$\approx \mathbf{0}$

Table 4.7: ANOVA and Tukey Test on Mean NET-PR among RATE-NUM Partitions (FilmTrust)

	DF	Sum. Sq.	Mean Sq.	F value	Pr (>F)
partition	2	0.0000774	3.869×10^{-5}	16.38	$\mathbf{1.06 \times 10^{-7}}$
Residuals	798	0.0018845	2.360×10^{-6}		

Pair	Diff. of Means	95% CI	p-value
B - A	0.000226	$[-8.6 \times 10^{-5}, 0.00053]$	0.20
C - B	0.000517	$[2.04 \times 10^{-4}, 0.00082]$	$\mathbf{3.2 \times 10^{-4}}$
C - A	0.000742	$[4.3 \times 10^{-4}, 0.00105]$	$\mathbf{1.0 \times 10^{-5}}$

Table 4.8: ANOVA and Tukey Test on Mean NET-PR among RATE-NUM Partitions (ciaoDVD)

	DF	Sum. Sq.	Mean Sq.	F value	Pr (>F)
partition	2	384331	192165	98.39	$\mathbf{2 \times 10^{-16}}$
Residuals	2928	5718398	1953		

Pair	Diff. of Means	95% CI	p-value
B - A	-1.4×10^{-6}	$[-3.3 \times 10^{-5}, 3.0 \times 10^{-4}]$	0.99
C - B	1.2×10^{-4}	$[9.1 \times 10^{-5}, 1.5 \times 10^{-4}]$	$\approx \mathbf{0}$
C - A	1.2×10^{-4}	$[8.9 \times 10^{-5}, 1.5 \times 10^{-4}]$	$\approx \mathbf{0}$

Does the mean NET-PR differ across RATE-NUM partitions? We repeat the previous setup, this time measuring popularity by means of PageRank. Tables 4.7 and 4.8 present the results on FilmTrust and CiaoDVD, respectively. The findings are similar, except with slightly lower significance: heaviness implies popularity.

Does the mean RATE-NUM differ across NET-PR partitions? Finally, we consider NET-PR partitions and study whether they contain users with significantly

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

Table 4.9: ANOVA and Tukey Test on Mean RATE-NUM among NET-PR Partitions (FilmTrust)

	DF	Sum. Sq.	Mean Sq.	F value	Pr (>F)
partition	2	17266	8633	12.73	$\mathbf{3.6 \times 10^{-6}}$
Residuals	801	543138	678		

Pair	Diff. of Means	95% CI	p-value
B - A	-0.239	$[-5.521, 5.043]$	0.99
C - B	9.948	$[4.666, 15.230]$	$\mathbf{3.3 \times 10^{-5}}$
C - A	9.709	$[4.427, 14.991]$	$\mathbf{5.3 \times 10^{-5}}$

Table 4.10: ANOVA and Tukey Test on Mean RATE-NUM among NET-PR Partitions (ciaoDVD)

	DF	Sum. Sq.	Mean Sq.	F value	Pr (>F)
partition	2	217992	108996	51.27	$\mathbf{2 \times 10^{-16}}$
Residuals	2739	5822399	2126		

Pair	Diff. of Means	95% CI	p-value
B - A	2.038	$[-3.02, 7.09]$	0.61
C - B	17.82	$[12.76, 22.88]$	$\approx \mathbf{0}$
C - A	19.86	$[14.80, 24.91]$	$\approx \mathbf{0}$

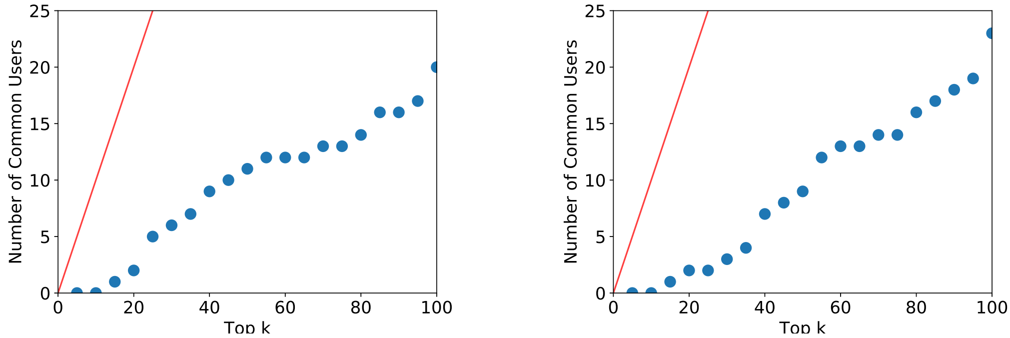
different RATE-NUM. Results are presented in Tables 4.9 and 4.10. As in the case of Degree partitions, popularity implies heaviness.

4.3 Ranking-Based Analysis

Based on the previous findings, we seek for further connections, this time among *very heavy raters* (top-100 users according to RATE-NUM) or *very popular users* (top-100 users according to NET-DEG or NET-PR). For FilmTrust that corresponds to about 13% of the users, while for CiaoDVD to about 4%.

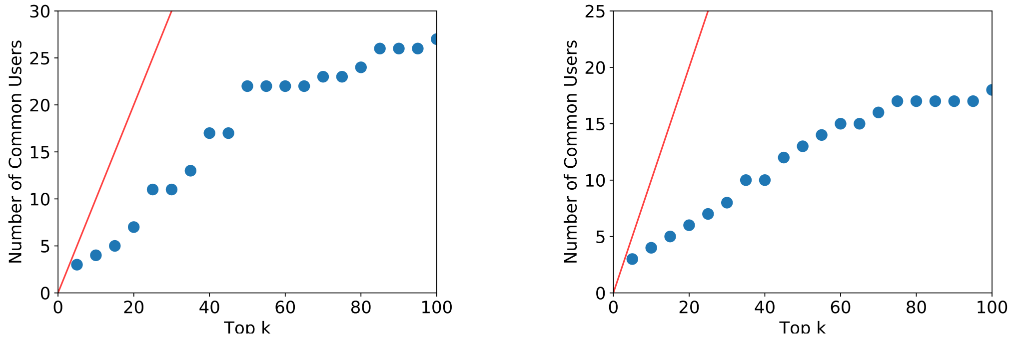
How many common users exist among the top-100 heavy and the top-100 popular? First we consider the number of common users across these rankings, with the results shown in Figures 4.3 and 4.4 for the two datasets. We see that the number of common users increases with a much lower rate than the maximum possible (drawn as the red line). Hence there exist more common users among the really heavy and the really popular users.

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



(a) RATE-NUM and NET-PR (b) RATE-NUM and NET-PR

Figure 4.4: Number of common users among the Top-K heaviest and most popular (NET-DEG, NET-PR) users (ciaoDVD)



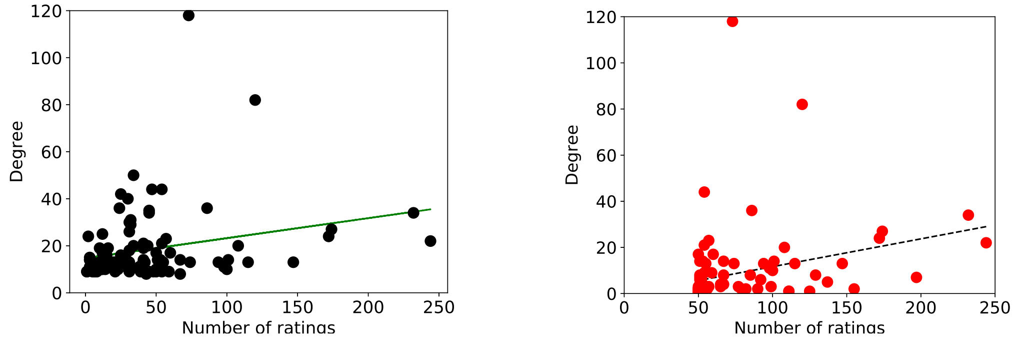
(a) RATE-NUM and NET-DEG (b) RATE-NUM and NET-PR

Figure 4.3: Number of common users among the Top-K heaviest and most popular (NET-DEG, NET-PR) users (FilmTrust)

Are RATE-NUM and NET-DEG correlated? We investigate whether heaviness and popularity (in terms of Degree) are correlated among the 100 most popular users or the 100 heaviest raters. For FilmTrust, Figure 4.5a shows the values of NET-DEG and RATE-NUM for each user among very popular users (according to NET-DEG), while Figure 4.5b shows the corresponding scatter plot for the very heavy raters. In both figures we draw the linear regression line, and also measure Pearson and Spearman's correlation coefficients. The very popular users have weak Pearson and Spearman correlation values of 0.25 and 0.27 with low significance (p-values of 0.01 and 0.07). In contrast, the very heavy users have weak Pearson but strong Spearman correlation values of 0.3 and 0.67 with high significance (p-values of 0.002 and ≈ 0).

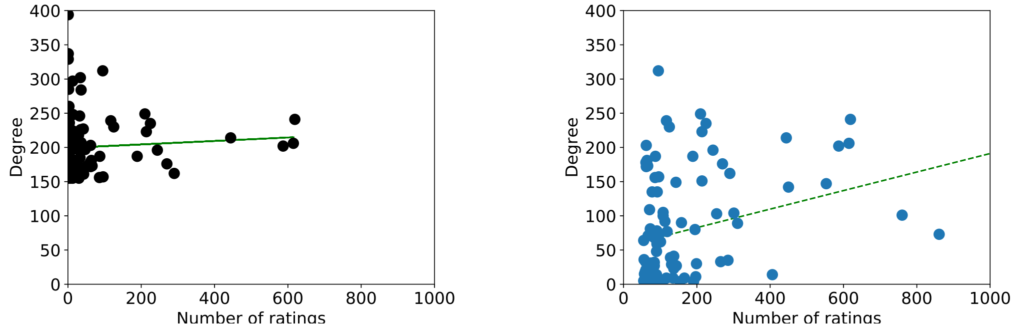
Similar results hold for the CiaoDVD dataset, shown in Figure 4.6. The very popular users exhibit non-significant weak correlation between heaviness and popularity, while the correlation in very heavy users is strong (Pearson and Spearman 0.31 and 0.44) and significant. These results imply that (1) overall there is a weak association between

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



(a) Very Popular Users (b) Very Heavy Raters

Figure 4.5: Scatter Plots (RATE-NUM, NET-DEG) (FilmTrust)



(a) Very Popular Users (b) Very Heavy Raters

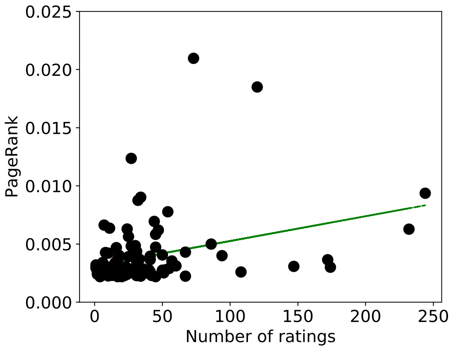
Figure 4.6: Plots (RATE-NUM, NET-DEG) (ciaoDVD)

RATE-NUM and NET-DEG in the very popular and the very heavy raters, and (2) RATE-NUM and NET-DEG are strongly correlated, in a non-linear sense, for the very heavy raters; the heavier the rater is, the more popular s/he becomes.

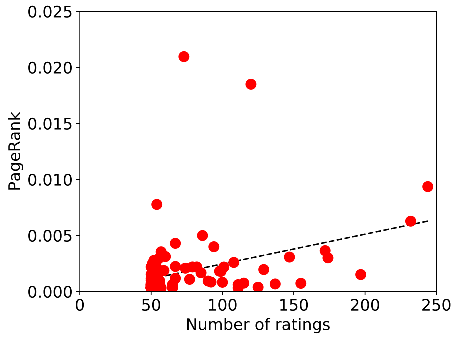
Are RATE-NUM and NET-PR correlated? We repeat the previous setup but this time define popularity by NET-PR. Figure 4.7 shows the results for FilmTrust, where the very popular users have an insignificant weak correlation among heaviness and popularity. On the other hand, the very heavy raters exhibit moderate to strong correlations (Pearson and Spearman 0.35 and 0.60) with high significance (p-values 0.004 and ≈ 0). Similar in CiaoDVD (scatter plots in Figure 4.8), heaviness and popularity among very heavy raters is moderately (Pearson and Spearman 0.37 and 0.45) correlated with high significance.

As a conclusion, we note that we have observed moderate to strong correlations among heavy users (top-100 by RATE-NUM) between their heaviness (RATE-NUM) and their popularity (NET-DEG and NET-PR). This correlation is not so much linear, as is rank-based (higher Spearman than Pearson correlation values).

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

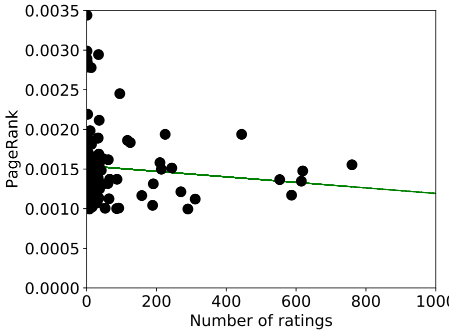


(a) Very Popular Users

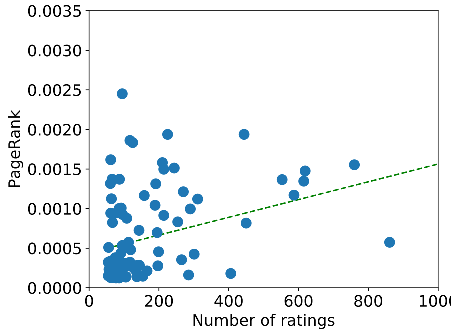


(b) Very Heavy Raters

Figure 4.7: Scatter Plots (RATE-NUM, NET-PR) (FilmTrust)



(a) Very Popular Users



(b) Very Heavy Raters

Figure 4.8: Scatter Plots (RATE-NUM, NET-PR) (ciaoDVD)

4.4 Conclusions

The work done in this chapter makes the first step towards studying the effects of social connections in rating behavior in social recommenders. At the level of users, we see that the number of ratings made by a user and her centrality in social network are related, particularly when the latter is measured in terms of the number of social connections. We have identified important strong connections between heaviness and popularity in social recommenders. In particular, the connection is stronger when we consider the very heavy raters, with strong evidence suggesting that heaviness implies popularity. In the next chapter, we build upon these results and examine pair of users in more detail. We consider pairs of highly popular and low popular users in the social network and examine if there is an observable influence in their rating behavior. We also investigate the opposite direction, i.e., whether friends with similar ratings have a leader-follower relationship in the social network then we extend our comparison to groups of users, extracted by community detection mechanisms or by ratings-based neighborhood approaches.

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

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

CHAPTER 5

Analysis at Pairwise Level

Pairwise level or Level 2 (L2) attributes quantify the *similarity* between pairs of users. For example, two users can have similar preferences in terms of ratings, or be socially similar in terms of their network distance or the number of common friends. Note that there are two separate ways to classify attributes: based on the view they correspond to (V1 or V2), and based on the level they are defined (L2). We pose two research questions based on our previous work findings in Chapter 4. At the second level (pairwise level), we begin with the research question that encompasses all pairs of users:

RQ2 Does high similarity between users in one view imply high similarity in the other?

The sub-question in this case is *Do Friends have similar ratings?* Similarity in terms of rating behavior (V1) quantifies how similar the ratings given by two users is. For this purpose, we use the widely popular adjusted cosine similarity [ELKR11] (RATE-SIM), which is related to Pearson's correlation coefficient. To address the case of implicit feedback, we measure the Jaccard index (RATE-JAC) of the sets of interactions between two users. For the second view, we consider various notions of network similarity, namely network distance (NET-DIST), the SimRank similarity[JW02] (NET-SIM), and the Leicht Holme Newman index[GZ16, LHN06] (NET-LHN). For this research question, we quantify relationships between the objects of study (user-pairs) between the two views using two methods. In the first, termed *partitioning*, we partition objects based on an attribute of one view (e.g., rating activity RATE-NUM), and investigate how the mean of an attribute of the other view (e.g., mean social activity in terms of NET-DEG) varies across partitions. This shows, for example, if degree centrality (NET-DEG) increases along with the number of ratings (RATE-NUM). In the second method, termed *ranking*, we compile two rankings of objects based on attributes of each view, and study whether statistical correlations between the rankings appear. For example, we may compute

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

how many pair of users are both highly similar (e.g., in the top-100) in terms of rating behavior (RATE-SIM) and in terms of social similarity (NET-LHN).

Next we focus on friends and pose the following research question:

RQ3 Does the level of user activity control the strength of friend similarity?

In other words, if we know individual aspects about users, e.g., their level of activity in a personalization system, can we infer a pairwise relationship, e.g., the similarity of their observed activities, between friends? for example, *Do popular users influence more?*

In this work, we consider pairs of friends and apply the following methodology. To answer questions such as when do two friends influence each other more, we classify friends into three groups, based on the amount of activity (rating or social) the two connected users exhibit, i.e., their node attributes. We consider pairs of friends that are: (LL) both of low activity, (HH) both of high activity, or (LH) one has high and the other low activity. We then investigate whether the rating/social similarities, i.e., the node attributes, differ significantly among the three groups.

5.1 Attributes

5.1.1 List of Attributes

Attributes are classified based on the views (V1 and V2). Attributes for the V1 (Ratings we refer to Figure 1.2) are:

RATE-SIM The adjusted cosine similarity alleviates problem of the the traditional cosine similarity where cosine similarity fails to consider the difference in rating scale among different users. The adjusted cosine similarity offsets this drawback by subtracting the corresponding user average from each co-rated pair.

RATE-JACC The Jaccard similarity index which is a popular measure for similarity between two sets of binary data. In our case, Jaccard similarity measure similarity between two users by considering the sets of items they have interacted with (implicit feedback).

RATE-PCC Pairwise similarity (RATE-PCC) is the rating similarity when only the common rated items between two users are considered.

Attributes for the V2 (Social Connections we refer to Figure 1.2) are:

NET-DIST Network distance between two users is a minimum number of connections, or links, that separate them in the network. It can also be defined as the length of the shortest path between two users.

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

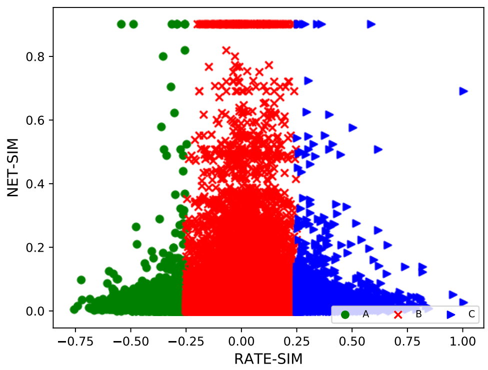
NET-SIM is SimRank.

NET-LHN is the *Leicht Holme Newman* index[GZ16, LHN06].

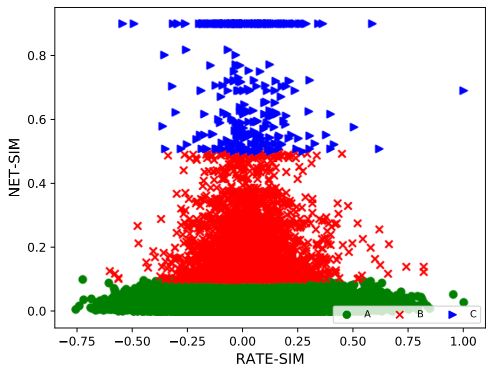
5.1.2 Distribution for All Pairs

Figure 5.1 shows scatter plots of RATE-SIM with each attribute (NET-DIST, NET-SIM, NET-LHN) of V2. We note that a point in these plots represents a pair of users. The color of a pair corresponds to the group (A, B, or C) it is partitioned into. Figure 5.2 shows scatter plots of RATE-JACC with each attribute (NET-DIST, NET-SIM) of V2.

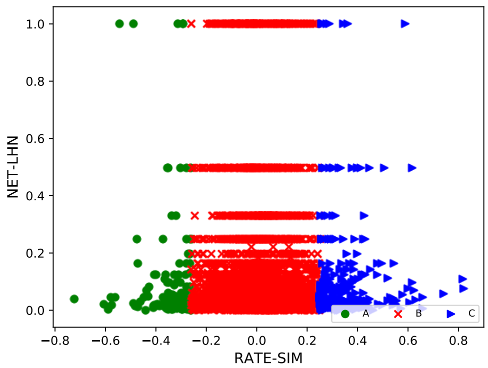
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.



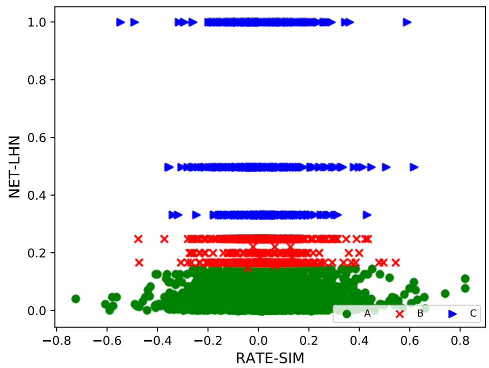
(a) Partitions on RATE-SIM



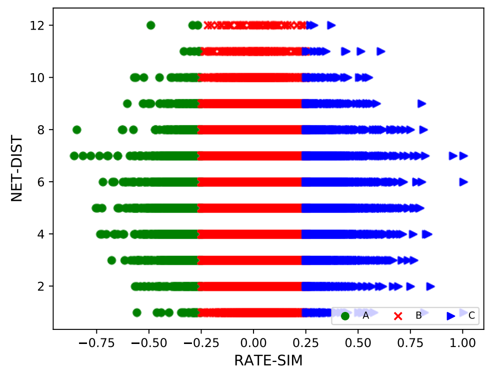
(b) Partitions on NET-SIM



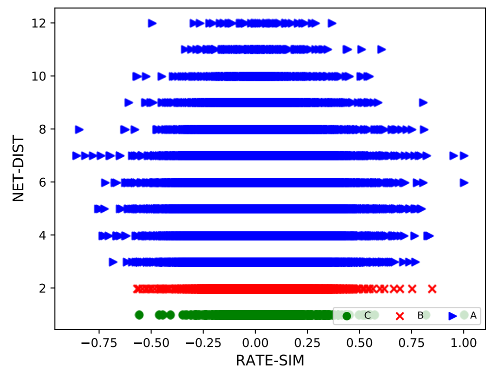
(c) Partitions on RATE-SIM



(d) Partitions on NET-LHN



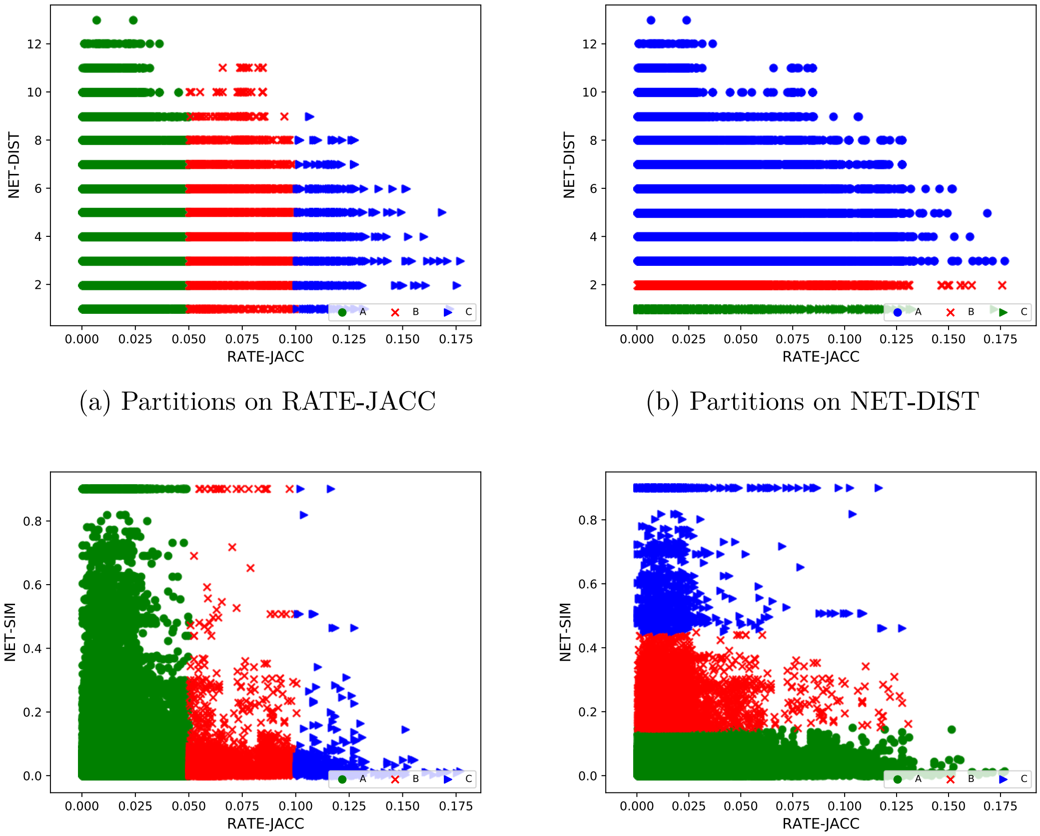
(e) Partitions on RATE-SIM



(f) Partitions on DIST-NET

Figure 5.1: Distribution of RATE-SIM values and Network similarity values among partitions (FilmTrust, All Pairs)

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



(a) Partitions on RATE-JACC

(b) Partitions on NET-DIST

(c) Partitions on RATE-JACC

(d) Partitions on NET-SIM

Figure 5.2: Distribution of RATE-JACC values and Network similarity values among partitions (FilmTrust, All Pairs)

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

5.1.3 Distribution for Friends

We compute the RATE-PCC and RATE-SIM for 1576 pairs of friends, and the mean RATE-PCC is 0.181 while the mean RATE-SIM is 0.049. For V2 we consider NET-SIM and NET-LHN with the mean value of 0.0186 and 0.0056. The distribution of RATE-SIM and RATE-PCC are shown in figures below.

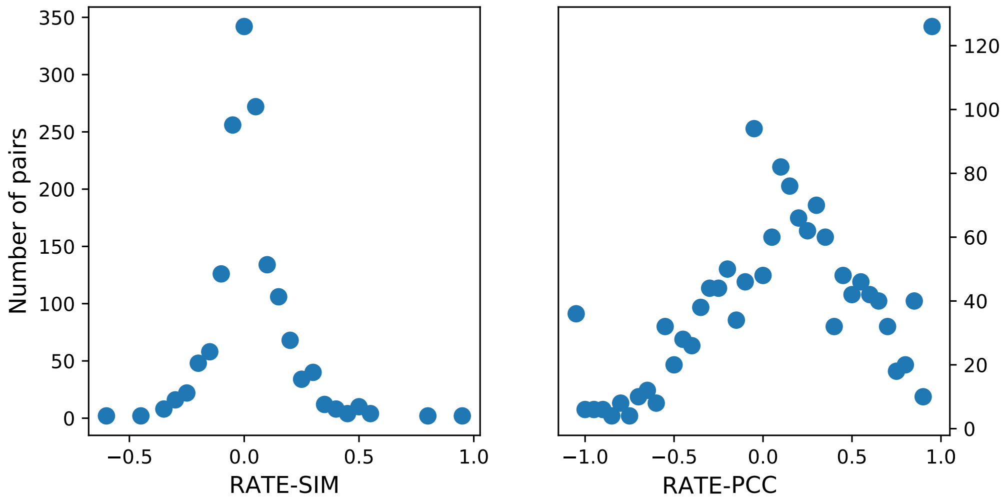


Figure 5.3: RATE-SIM and RATE-PCC probabilty distribution (FilmTrust, Friends)

5.2 Analysis for All Pairs

5.2.1 Results from Partition-Based Analysis

We partition pairs of users based on four attributes, three on network similarity, NET-DIST, NET-SIM, NET-LHN, one on rating similarity, RATE-SIM and RATE-JACC. Each division splits pairs into three groups, A, B, C, with C having the most similar pairs of users (in the case of NET-DIST, this means pairs with the lowest NET-DIST). Partitions are unbalanced, and shown in Table 5.1.

Table 5.1: Description of Partitions for Similarity between Pairs of Users

	upper div	lower div
NET-DIST	2	3
NET-SIM	0.5	0.1
NET-LHN	0.25	0.15
RATE-JACC	0.1	0.05
RATE-SIM	0.25	-0.25

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

Does the mean RATE-SIM differ across NET-DIST partitions? In this experiment, we partition pairs of users according to their distance. In total there are 222,103 number of pairs considered; note that we exclude pairs with RATE-SIM zero. Partition C contains 1,866 pairs of friends; partition B contains 10,842 pairs of users with NET-DIST of 2 (friends of friends); partition A contains 209,395 pairs of users with NET-DIST greater than 2. Partition C, which contains pairs of friends, has the highest mean RATE-SIM of 0.041. In partition B, the mean RATE-SIM drops to 0.022, while among all other pairs, in partition A, the mean RATE-SIM is 0.020. Therefore, we observe an increase in the mean RATE-SIM as the network distance decreases, a phenomenon which we investigate. ANOVA shows that the mean RATE-SIM across NET-DIST partitions changes significantly (p-value in the order of 10^{-16}). Following this finding, we perform Dunnett's T3 test to check the significance of the pairwise differences between means; Table 5.2 presents the pairwise mean differences and their 95% confidence intervals. We observe that all differences are significant (the intervals do not contain the null hypothesis value of zero), with partition C having the largest mean RATE-SIM over the other partitions.

Table 5.2: Dunnett's T3 Test on Mean RATE-SIM among NET-DIST Partitions

Pair	Diff. of Means	95% CI
C - B	0.0184	$[0.0174, 0.0193]$
C - A	0.0203	$[0.0193, 0.0214]$
B - A	0.0020	$[0.0010, 0.0029]$

Does the mean RATE-SIM differ across NET-SIM partitions? We partition pairs of users according to their NET-SIM. We have 204,608 number of pairs in total; note that we exclude pairs where RATE-SIM or NET-SIM is zero. Group C contains 752 pairs of users with the high similarity values; group B has 5,836 pairs of users with the medium NET-SIM; and group A contains 198,020 pairs of users with the lowest NET-SIM. In each partition, we compute the mean RATE-SIM, and observe that the means are different across the groups. ANOVA shows that the mean RATE-SIM across NET-SIM partitions changes significantly (p-value of 2×10^{-16}). Group C has the most similar pairs of users with a mean of 0.024; in group B the mean is 0.022, which is greater than the mean 0.018 of the group A. We thus observe moderate differences between the groups. Dunnett's T3 test shows that these differences are also significant, as shown in Table 5.3.

Does the mean RATE-SIM differ across NET-LHN partitions? We partition pairs of users according to their NET-LHN. We have 10,932 number of pairs in total; note that we exclude pairs where RATE-SIM or NET-LHN is zero. Group C contains 1,124 pairs of users; Group B has 1,030 pairs of users with the medium NET-LHN and the last Group A contains 8,778 pairs of users with the lowest NET-LHN values. The mean RATE-SIM is computed in each group, and we see that the means are roughly equal; A and C have mean RATE-SIM of 0.022, while B has mean RATE-SIM of 0.023.

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

Table 5.3: Dunnett's T3 Test on Mean RATE-SIM among NET-SIM Partitions

Pair	Diff. of Means	95% CI
B - A	0.0035	[0.0024, 0.0046]
C - A	0.0061	[0.0051, 0.0072]
C - B	0.0026	[0.0015, 0.0037]

The ANOVA verifies that differences are not significant, and thus we perform no post hoc test.

For the next three experiments, we partition pairs of users according to their rating similarity (RATE-SIM).

Does the mean NET-DIST differ across RATE-SIM partitions? RATE-SIM partitions encompasses 222,103 number of pairs in total. Group C contains 12,768 pairs of users; group B has 203,051 pairs of users and group A contains 6,284 pairs of users with the lowest RATE-SIM values. The mean NET-DIST is computed in each group. Group A which contains most dissimilar pairs of users has the highest mean of 5.25, group B with neither similar nor dissimilar pairs has the mean of 4.93, and group C which contains the most similar users has the mean of 5.16. The results of ANOVA shows a significant difference of means across partitions with a p-value of 2×10^{-16}.

Dunnett's T3 test, depicted on Table 5.4, shows that all pairwise differences are significant. We observe some moderate variation in the magnitude of the differences. Specifically, groups A and C (of strong similarity or dissimilarity) have the highest mean NET-DIST of 5.25 and 5.16, compared to 4.94 of group B. This implies that pairs of users that are moderately similar (group B) tend to be somewhat closer in terms of network distance. The most important result however is negative. Looking at highly similar raters, we find no relationship about their network position: they can either be directly connected or very far from each other. This is in contrast to the opposite direction of correlation between RATE-SIM and NET-DIST (Table 5.2), where direct connection of users implies higher similarity in rankings.

Table 5.4: Dunnett's T3 Test on Mean NET-DIST among RATE-SIM Partitions

Pair	Diff. of Means	95% CI
B - A	-0.3098	[-0.3224,-0.2971]
C - A	-0.0876	[-0.1003,-0.07501]
C - B	0.2221	[0.2094,0.2348]

Does the mean NET-SIM differ across RATE-SIM partitions? We conduct the same experiment on RATE-SIM partitions by exploring the mean NET-SIM. There are 204,608 number of pairs examined in total, with group C containing 11,676 pairs, group B with 186,724 pairs, and group A with 6,208 pairs. Mean NET-SIM is roughly equal

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

across groups: B has the highest mean of 0.032, followed by A with 0.031, and C with 0.030. Results of ANOVA show statistical significance (p-value of 2.64×10^{-13}), and post hoc analysis results are shown in Table 5.5. The differences are not always significant, and their strength is very small. This leads us to the conclusion that the three groups do not differ.

Table 5.5: Dunnett's T3 Test on Mean NET-SIM among RATE-SIM Partitions

Pair	Diff. of Means	95% CI
B - A	0.0004	$[-8.03 \times 10^{-6}, 0.0008]$
C - A	-0.0010	$[-1.47 \times 10^{-3}, -0.0006]$
C - B	-0.0014	$[-1.84 \times 10^{-3}, -0.0010]$

Does the mean NET-LHN differ across RATE-SIM partitions? The last experiment is based on 10,932 pairs of users, with group A having 604 pairs, group B having 9,376 pairs, and group C having 952 pairs. The mean NET-LHN in these partitions are roughly equal, with values 0.10, 0.12, 0.10, respectively. Although ANOVA sees significant differences (p-value of 1.73×10^{-15}), as well as Dunnett's T3 test (Table 5.6), the differences of means are generally small and not indicative of correlation.

Table 5.6: Dunnett's T3 Test on Mean NET-LHN among RATE-SIM Partitions

Pair	Diff. of Means	95% CI
B - A	0.0189	[0.0123,0.0254]
C - A	0.0064	[0.0001,0.0126]
C - B	-0.0125	[-0.0189,-0.0061]

Does the mean RATE-JACC differ across NET-DIST partitions? We partition pairs of users according to their NET-DIST. We have 279,640 number of pairs in total; where zero values are excluded from both RATE-JACC or NET-DIST . Group C contains 2236 pairs of friends; Group B has 13775 pairs of friends of friends and the last Group A contains 263,629 pairs of users with the NET-DIST greater then 2. The mean RATE-JACC is computed in each group, and we see that the means that as the mean increases as the distance decreases. Partition C of friends has the high RATE-JACC mean of 0.028, B partition has the mean of 0.026 , while A has the lowest mean RATE-JACC of 0.019. The ANOVA shows that differences of the mean RATE-JACC across NET-DIST partitions changes significantly (p-value of 2×10^{-16}). and thus we perform post hoc test. Results are shown in Table 5.7.

Does the mean RATE-JACC differ across NET-SIM partitions? We partition pairs of users according to their NET-SIM. We have 316,970 number of pairs in total; note that we exclude pairs where RATE-JACC or NET-SIM is zero. Group C contains 1398 pairs of users with the high NET-SIM similarity values; group B has 4641 pairs of users with the medium NET-SIM; and group A contains 310,931 pairs of users with the lowest NET-SIM. In each partition, we compute the mean RATE-JACC, and observe that

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

Table 5.7: Dunnett's T3 Test on Mean RATE-JACC among NET-DIST Partitions

Pair	Diff. of Means	95% CI
C - B	0.001	[0.0001, 0.0020]
C - A	0.009	[0.0077, 0.0103]
B - A	0.008	[0.0075, 0.0085]

the means are different across the groups. Group C has the most similar pairs of users has the high mean RATE-JACC of 0.018; in group B the mean is 0.017, which is slightly greater than the mean 0.016 of the group A. We thus observe moderate differences among groups. Dunnett's T3 test shows that these differences are also significant, as shown in Table 5.8.

Table 5.8: Dunnett's T3 Test on Mean RATE-JACC among NET-SIM Partitions

Pair	Diff. of Means	95% CI
C - B	0.0002	[-0.0009, 0.0013]
C - A	0.0005	[-0.0005, 0.0014]
B - A	0.0003	[-0.0003, 0.0007]

Does the mean NET-DIST differ across RATE-JACC partitions? We conduct the same experiment on RATE-JACC partitions by exploring the mean NET-DIST. There are 279,640 number of pairs examined in total, with group C containing 2740 pairs of most similar users, group B with 14079 pairs, and group A with 262821 pairs. Mean NET-DIST varies across groups: A has the highest mean NET-DIST of 5, followed by B with 4.4, and C with 3.6. Results of ANOVA show statistical significance (p-value of 2×10^{-16}), and post hoc analysis results are shown in Table 5.9. The differences are significant, This leads us to the conclusion that the three groups differ, Group C which contains most similar users has a small mean NET-DIST, it implies that most similar users tend to be socially close (in the network)

Table 5.9: Dunnett's T3 Test on Mean NET-DIST among RATE-JACC Partitions

Pair	Diff. of Means	95% CI
C - B	-0.5634	[-0.6478,-0.4791]
C - A	-1.1357	[-1.2133,- 1.0581]
B - A	-0.5722	[-0.6072,-0.5373]

Does the mean NET-SIM differ across RATE-JACC partitions? We partition on RATE-JACC by exploring the mean NET-SIM. There are 316,970 number of pairs in total, with group C containing 2,091 pairs, group B with 9,464 pairs, and group A

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

with 305,415 pairs. Mean NET-SIM is roughly equal for group B and group A. B has the mean of 0.032 and the mean NET-SIM for A is 0.031. C has the lowest mean of 0.021. Results of ANOVA show statistical significance with (p-value of 2×10^{-16}) and post hoc analysis results shown in Table 5.10. The differences are not significant. As conclusion the three groups do not differ.

Table 5.10: Dunnett's T3 Test on Mean NET-SIM among RATE-JACC Partitions

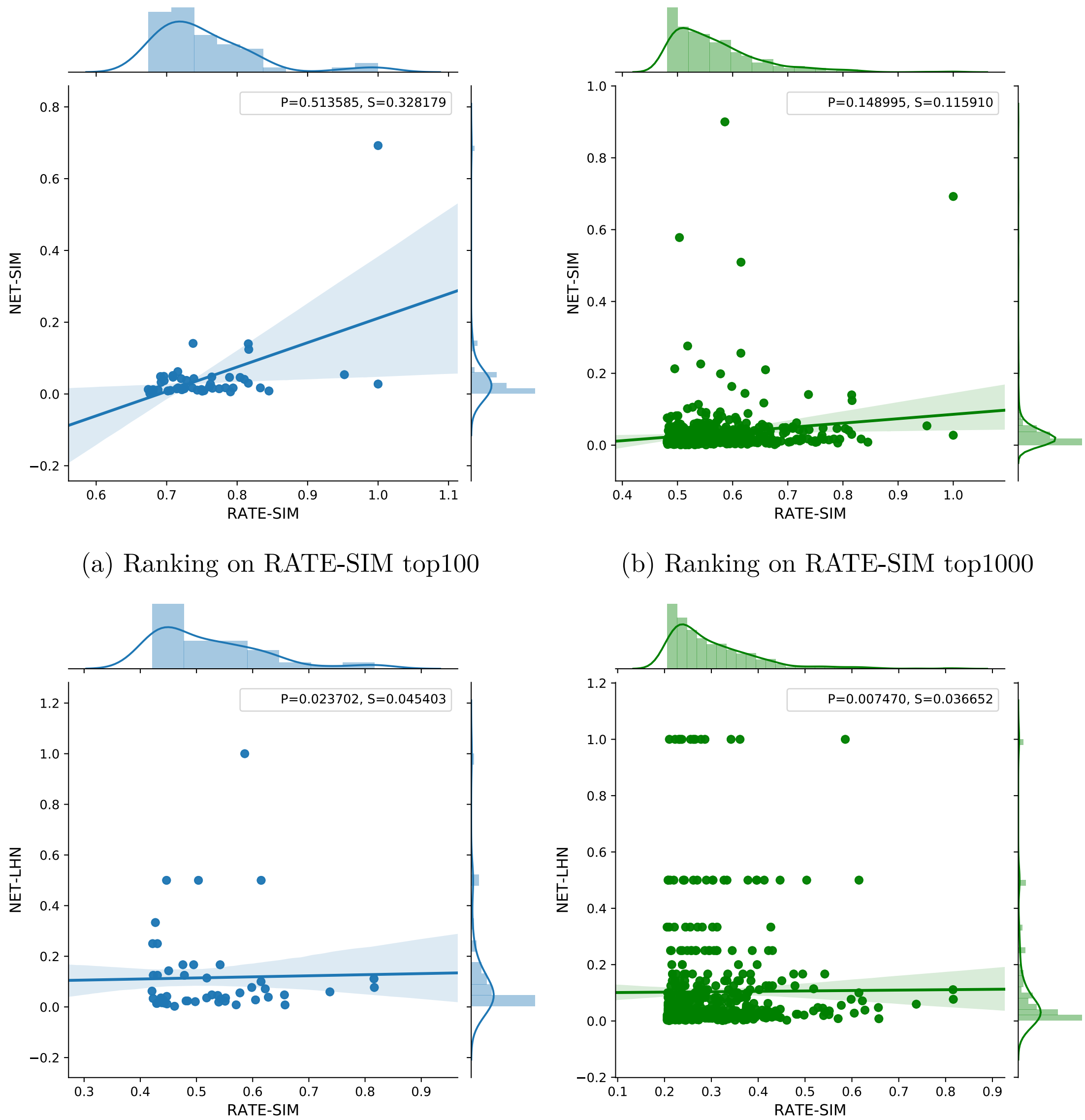
Pair	Diff. of Means	95% CI
C - B	-0.0120	[-0.0152,-0.0088]
C - A	-0.0101	[-0.0130,-0.0072]
B - A	0.0018	[0.0005,0.0032]

5.2.2 Results from Ranking-Based Analysis

We next present results by looking at the rankings created by attributes of similarity among pairs of users. For the first set of experiments, we look at ratings (top-100 and top-1000) based on RATE-SIM, and look for correlations between RATE-SIM and one of the social similarity attributes.

Are RATE-SIM and NET-SIM correlated for top RATE-SIM pairs? We observe a moderate correlation for the top-100 RATE-SIM pairs (Figure 5.4a). The correlation, however weakens as we look at the top-1000 pairs (Figure 5.4b).

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



(a) Ranking on RATE-SIM top100

(b) Ranking on RATE-SIM top1000

(c) Ranking on RATE-SIM top100

(d) Ranking on RATE-SIM top1000

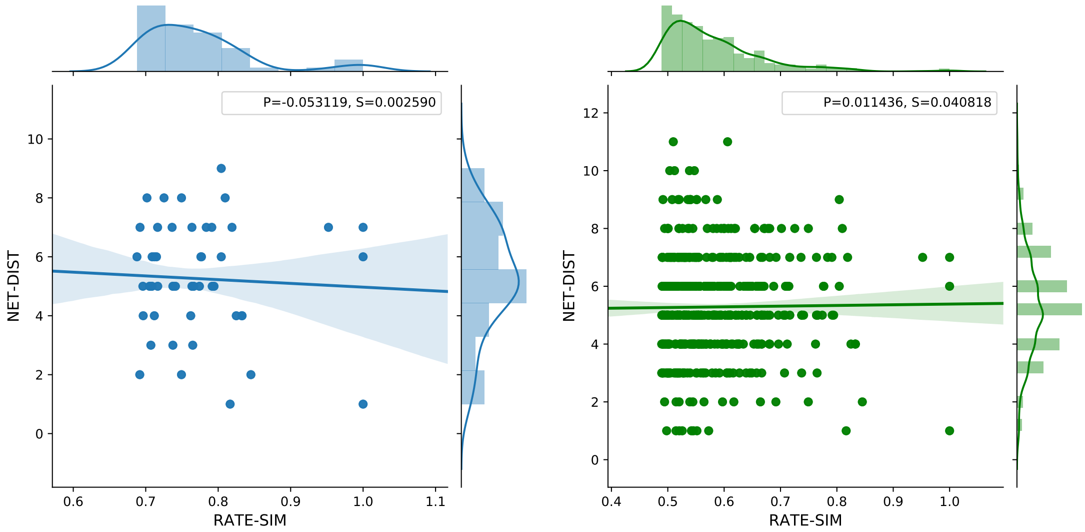
Figure 5.4: Ranking correlation results (FilmTrust, All Pairs)

Are RATE-SIM and NET-LHN correlated for top RATE-SIM pairs? A weak correlation between RATE-SIM and NET-LHN is observed for top-100 most similar pairs by rating (RATE-SIM) on Figure 5.4c. Again the correlation significantly weakens as we increase the number of examined pairs to 1000 (Figure 5.4d).

Are RATE-SIM and NET-DIST correlated for top RATE-SIM pairs? For the top-100 similar pairs, we observe a very weak positive correlation between network

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

similarity in terms of NET-DIST (recall than NET-DIST is a measure of dis-similarity) and RATE-SIM. As we increase the number of pairs to 1000, the correlations weaken.



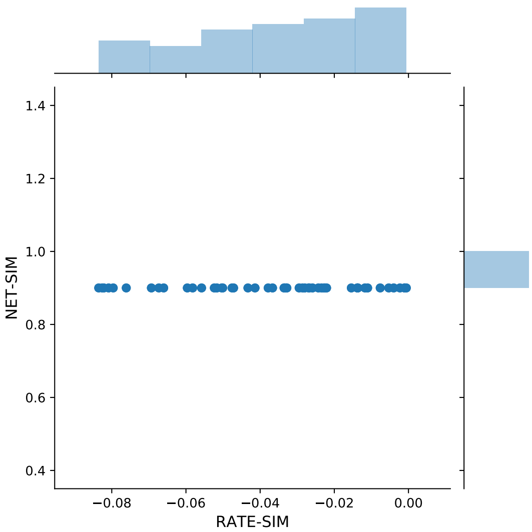
(a) Ranking on RATE-SIM top100 (b) Ranking on RATE-SIM top1000

Figure 5.5: Ranking correlation results (RATE-SIM and NET-DIST correlation for top RATE-SIM pairs)(FilmTrust, All Pairs)

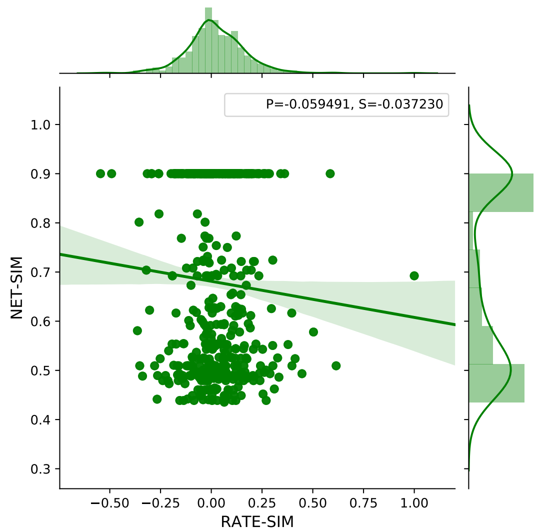
In the last set of experiments, we look at three rankings induced by an attribute measuring social similarity, namely by NET-SIM, NET-LHN, and NET-DIST.

Are NET-SIM and RATE-SIM correlated for top NET-SIM pairs? All top-100 pairs have NET-SIM of 0.9, which means we cannot compute correlations between the tested attributes (Figure 5.6a). When we increase the number of examined pairs to 1000, we observe very weak correlations (Figure 5.6b).

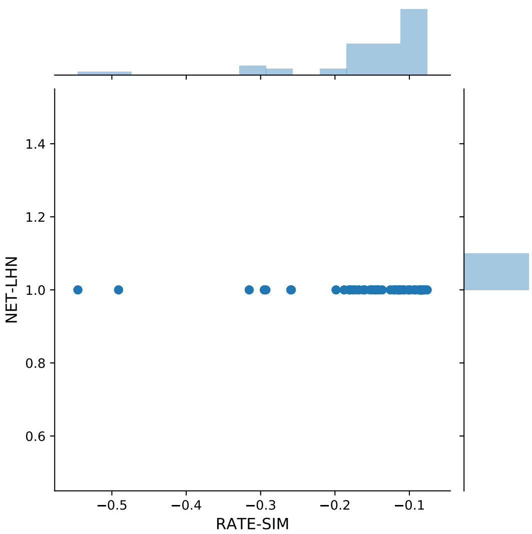
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.



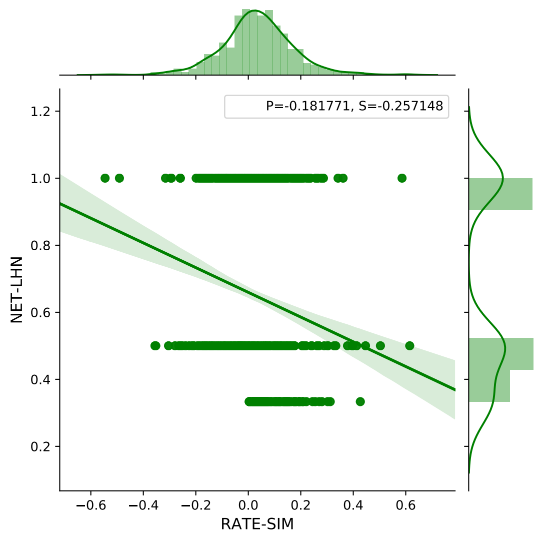
(a) Ranking on NET-SIM top100



(b) Ranking on NET-SIM top1000



(c) Ranking on NET-LHN top100



(d) Ranking on NET-LHN top1000

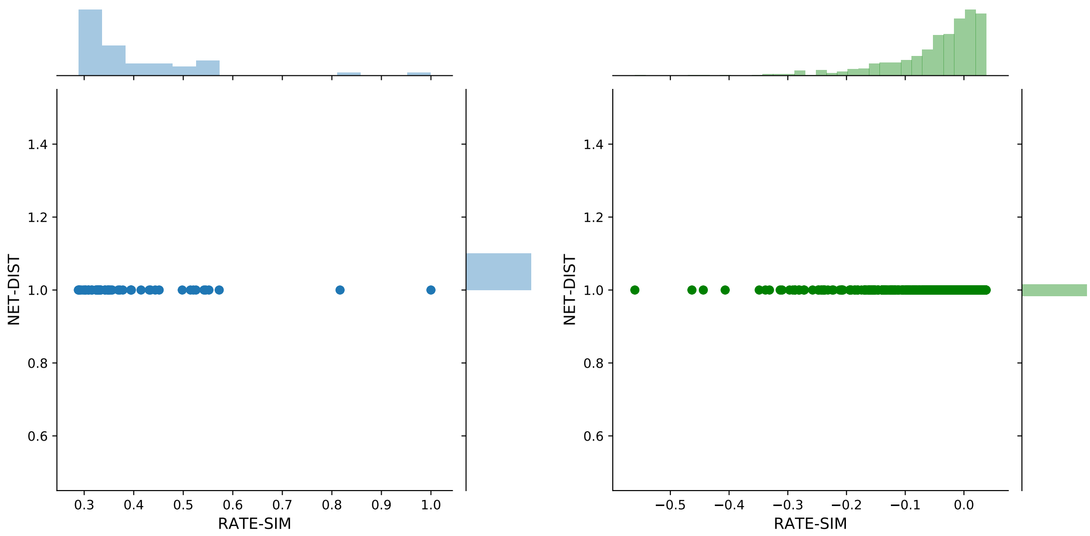
Figure 5.6: Ranking correlation results (FilmTrust, All Pairs)

Are NET-LHN and RATE-SIM correlated for top NET-LHN pairs? As before, looking at the top-100 pairs by NET-LHN, we cannot draw any conclusions (Figure 5.6c), as all pairs have the highest NET-LHN value of 1. Increasing the number of pairs to 1000, we begin to see weak negative correlations between NET-LHN and RATE-SIM (Figure 5.6d), implying that higher RATE-SIM is related to lower NET-LHN.

Are NET-DIST and RATE-SIM correlated for top NET-DIST pairs? We look

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

at the top-100 and top-1000 pairs that have the lowest NET-DIST, respectively in Figures 5.7a and 5.7b. In both cases, this means pairs of friends with distance of 1. As a result, we cannot draw conclusions on correlation between NET-DIST and RATE-SIM using the ranking method.



(a) Ranking on NET-DIST top100 (b) Ranking on NET-DIST top1000

Figure 5.7: Ranking correlation results (NET-DIST and RATE-SIM correlation for top NET-DIST pairs) (FilmTrust, All Pairs)

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

5.3 Analysis for Friends

5.3.1 Methodogy

Previous work that exploits social influence between users [RRS15, AV16, MSW18] has demonstrated that there exist correlations between the similarities in terms of the social network and the observed feedback. In terms of our augmented social network, this translates into correlations of the various edge attributes. In this work, we seek to understand when these correlations are stronger. Specifically, we want to see if node attributes can help identify these instances.

Therefore, we define classes of pairs of friends, based on their node attributes, and then measure whether similarities among edge attributes become stronger or weaker across classes. More concretely, a user is assigned a label L when her activity (node attribute RATE-NUM or NET-DEG) is below some threshold L, label H when her activity is above another threshold H, and no label otherwise; we consider various values for these thresholds. In this way, two friends are classified into four classes:

LL when both have label L,

HH when both have label H,

LH when one has label L and the other label H,

– when one has no label.

This essentially induces a partition on the edges of the augmented social network. We examine the three classes LL, HH, and LH, to see if for some class we measure stronger/weaker edge-based similarities. As a first step, we plot the distribution of an edge attribute (RATE-SIM, RATE-PCC, NET-SIM, NET-LHN) within the class, and visually explore if any differences across classes appear. Then, we focus on the mean edge attribute for a class, and perform statistical tests (ANOVA followed by pairwise post hoc analysis) to see whether the visual differences across classes are actually significant. In total, there are 740 users with 1576 social connections. Across all users, mean NET-DEG is 18, and mean RATE-NUM is 43.5. Across all pairs of friends, mean RATE-PCC is 0.181, mean RATE-SIM is 0.049, mean NET-SIM is 0.0186, while mean NET-LHN is 0.0056.

The value of the attribute that determines the class (L or H) can vary. For NET-DEG, the following boundary values are considered: 5, 10, 15, 20, 30, 40, and 50. For example, class L is defined as NET-DEG $\leq$ 15 and class H as NET-DEG $\geq$ 30 and for RATE-NUM, the following boundary values are considered: 5, 10, 20, 30, 50, 70, and 100. For example, class L is defined as RATE-NUM $\leq$ 30 and class H as RATE-NUM $\geq$ 50. We create a 7x1 table that computes the average RATE-PCC of the pairs in the HH pair class for each boundary value defining H and a 7x1 table that computes the average RATE-PCC of the pairs in the LL pair class for each boundary value defining L. At the end we create

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

a 7x7 table that computes the average RATE-PCC of the pairs in the LH pair class for each boundary value defining L and for each boundary value defining H. For example, the Table 5.11 entry (L, H) for boundaries (10, 20) shows the RATE-PCC in pairs when one user has NET-DEG ≤ 10 and the other NET-DEG ≥ 20.

Visualizing Distribution and Probability Density of the Groups

Table 5.11: Mean RATE-PCC based on NET-DEG boundary values

		H	5	10	15	20	30	40	50
		HH	0.18	0.162	0.166	0.162	0.121	0.27	-0.17
L	LL	LH							
5	0.152		0.188	0.212	0.28	0.275	0.27	0.27	0.193
10	0.153		0.201	0.235	0.28	0.29	0.27	0.26	0.201
15	0.132		0.185	0.204	0.258	0.265	0.258	0.27	0.164
20	0.14		0.19	0.201	0.251	0.257	0.25	0.248	0.157
30	0.16		0.192	0.201	0.248	0.255	0.246	0.246	0.154
40	0.17		0.19	0.195	0.23	0.233	0.22	0.233	0.15
50	0.18		0.192	0.2	0.227	0.23	0.214	0.232	0.151

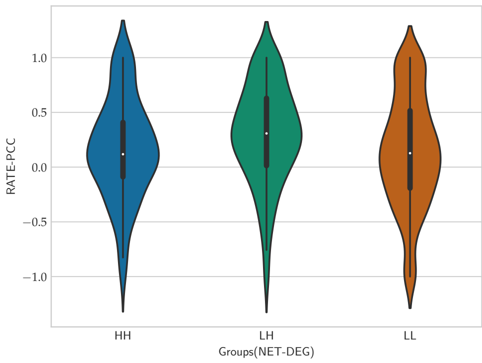
From the 7x7 table that computes the average RATE-PCC for pair class LH, we pick one average PCC and consider the L Class and H Class in that boundary; ANOVA is used to check how the mean varies ac cross the three classes. All details are shown in Table 5.11; from this table we observe three different means that are highlighted in the green color; the one above is one of HH pair-class, on the left side we have the mean RATE-PCC of LL pair-class and at the upper middle location of LH pair-class we have the mean RATE-PCC, we use those boundaries to make three different groups and analyze the rating behavior of friends that are categorized in these three different pair-classes (Groups).

5.3.2 Results from Partition-Based analysis

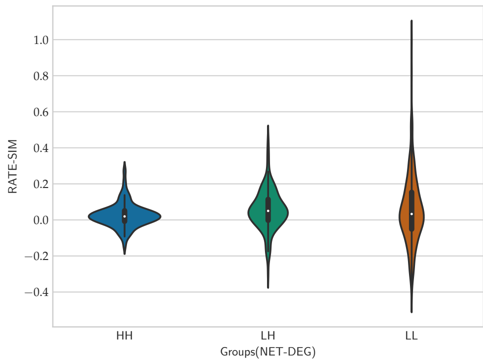
Does RATE-PCC depend on NET-DEG? We first consider partitioning pairs of friends based on the NET-DEG. We explore different definition of low (L) and high (H) NET-DEG, based on which we assign pairs of friends into classes LL, LH, and HH. For each class, we compute the mean RATE-PCC. The results are shown in Table 5.12, where we see that RATE-PCC varies significantly across different classes.

We then fix L and H to their default values of L=10 and H=20, and look deeper into the three classes they induce. Specifically, LL contains pairs of friends where each has less than 10 friends in total; HH contains pairs of friends where each has more than 20 friends in total; LH contains pairs of friends, where one has few (≤ 10) and the other has many (≥ 20) other friends. There are 873 number of pairs examined in total; HH contains 142 pairs, LH has 157 pairs, and LL 574 pairs. The mean RATE-PCC within the classes is 0.162, 0.293 and 0.137 respectively.

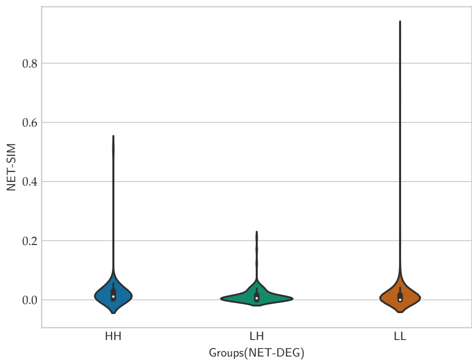
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub



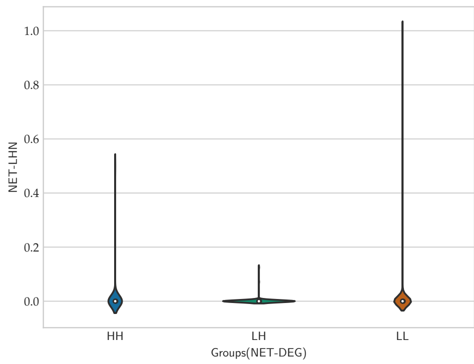
(a) RATE-PCC vs. NET-DEG



(b) RATE-SIM vs. NET-DEG



(c) NET-SIM vs. NET-DEG



(d) NET-LHN vs. NET-DEG

Figure 5.8: Classes based on NET-DEG

Table 5.12: Mean RATE-PCC of NET-DEG classes

		H	5	10	15	20	30	40	50
		HH	0.18	0.162	0.166	0.162	0.121	0.27	-0.17
L	LL	LH							
5	0.152		0.188	0.212	0.28	0.275	0.27	0.27	0.193
10	0.153		0.201	0.235	0.28	0.29	0.27	0.26	0.201
15	0.132		0.185	0.204	0.258	0.265	0.258	0.27	0.164
20	0.14		0.19	0.201	0.251	0.257	0.25	0.248	0.157
30	0.16		0.192	0.201	0.248	0.255	0.246	0.246	0.154
40	0.17		0.19	0.195	0.23	0.233	0.22	0.233	0.15
50	0.18		0.192	0.2	0.227	0.23	0.214	0.232	0.151

Figure 5.8a shows the distribution of RATE-PCC between pairs of friends in each of the three classes. While not immediately apparent, the distributions have different means and shape. To quantify this, we perform ANOVA analysis, which shows that the mean

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

RATE-PCC across the classes is significantly different (p-value of 0.00235). Then, post hoc analysis of the results, presented in Table 5.13, finds that the RATE-PCC similarity of LH pairs of friends is considerably and significantly higher than other pairs of friends. This implies that a pair of friends that is formed by a popular H user and a less popular L user tend to influence each other's rating behavior.

Table 5.13: RATE-PCC differences across NET-DEG classes

Pair	Diff. of Means	95% CI
LL - LH	−0.1551	[−0.254, −0.0562]
LL - HH	−0.0253	[−0.0129,0.0786]
LH - HH	0.1297	[0.0056,0.2539]

Does RATE-SIM depend on NET-DEG? We repeat the previous setup, this time looking at the RATE-SIM between two friends. Table 5.14 shows the mean RATE-SIM for various definitions of L and H in terms of NET-DEG. Differences exists but are not as dramatic as in the case of RATE-PCC.

Table 5.14: Mean RATE-SIM of NET-DEG classes

		H	5	10	15	20	30	40	50
		HH	0.05	0.037	0.03	0.025	0.025	0.019	-0.013
L	LL	LH							
5	0.06		0.05	0.042	0.06	0.06	0.06	0.07	0.039
10	0.05		0.05	0.05	0.05	0.05	0.05	0.05	0.027
15	0.05		0.05	0.05	0.06	0.06	0.06	0.06	0.026
20	0.05		0.05	0.05	0.06	0.06	0.05	0.05	0.026
30	0.05		0.05	0.044	0.05	0.05	0.05	0.05	0.023
40	0.05		0.05	0.044	0.05	0.05	0.05	0.05	0.023
50	0.05		0.05	0.044	0.05	0.05	0.04	0.05	0.023

Fixing the definition of L and H to their default values, in Figure 5.8b, we plot the distribution of RATE-SIM within the three classes. Class HH has a mean RATE-SIM of 0.025, LH of 0.05, and LL of 0.05. That is, mean RATE-SIM is roughly equal for LH and LL categories and higher than HH which has the lowest mean. However, ANOVA results show that the differences are not significant (p-value of 0.148). We conclude that no safe conclusions can be drawn from this experiment.

Does NET-SIM depend on NET-DEG? In this experiment we measure friend similarity in terms of their global network similarity quantified as NET-SIM. Table 5.15 presents the mean NET-SIM for the various classes previously explored, where we do not observe any meaningful trends.

We next fix L and H to their default values, and plot the distribution of NET-SIM within the three induced classes in Figure 5.8c. Classes LL and HH have a mean of 0.02, while LH has a mean of 0.014, i.e., they are roughly equal. ANOVA finds they do

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

Table 5.15: Mean NET-SIM of NET-DEG classes

		H	5	10	15	20	30	40	50
		HH	0.02	0.02	0.02	0.02	0.016	0.013	0.002
L	LL	LH							
5	0.03		0.013	0.014	0.014	0.014	0.01	0.01	0.01
10	0.02		0.015	0.016	0.014	0.014	0.012	0.01	0.01
15	0.02		0.016	0.018	0.017	0.017	0.016	0.014	0.01
20	0.02		0.016	0.019	0.018	0.017	0.017	0.014	0.01
30	0.02		0.018	0.02	0.02	0.02	0.018	0.014	0.01
40	0.02		0.018	0.02	0.02	0.02	0.017	0.015	0.01
50	0.02		0.018	0.02	0.02	0.02	0.017	0.014	0.01

not significantly differ (p-value of 0.466). Any differences in terms of NET-SIM across NET-DEG classes are not significant.

Does NET-LHN depend on NET-DEG? In the last experiment with classes defined on NET-DEG, we measure pairwise similarities in terms of the local network similarity NET-LHN. Table 5.16 presents the mean NET-LHN for the studied classes.

Table 5.16: Mean NET-LHN of NET-DEG classes

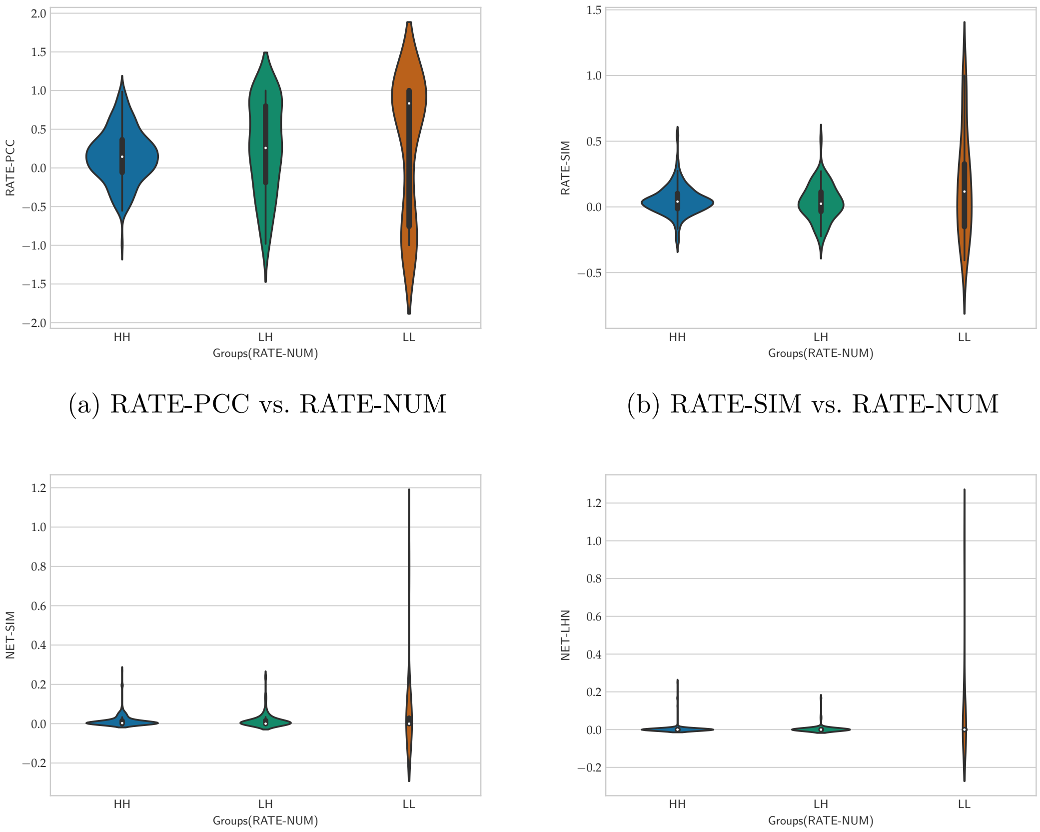
		H	5	10	15	20	30	40	50
		HH	0.005	0.007	0.008	0.007	0	0	0
L	LL	LH							
5	0.011		0.001	0.001	0.0007	0.0008	0	0	0
10	0.006		0.004	0.005	0.001	0.001	0.001	0	0
15	0.007		0.004	0.005	0.003	0.003	0.003	0.002	0
20	0.007		0.004	0.006	0.004	0.003	0.004	0.002	0
30	0.007		0.005	0.007	0.006	0.006	0.004	0.001	0
40	0.007		0.005	0.006	0.006	0.006	0.003	0.001	0
50	0.007		0.005	0.006	0.006	0.005	0.003	0.001	0

We fix L and H to their default values for NET-DEG, and drawn the distribution of NET-LHN across the three classes in Figure 5.8d. ANOVA reports no significant differences for the mean NET-DEG values.

Does RATE-PCC depend on RATE-NUM? In the following set of experiments, we classify pair of friends based on their number of provided ratings, RATE-NUM. First, we consider pairwise similarity in terms of RATE-PCC. Table 5.17 includes the mean RATE-NUM for different definitions of L and H in terms of RATE-NUM. Except when L=5, we note that the mean RATE-PCC is roughly the same across classes.

We fix L and H to their default values L=10 and H=30, and examine the three classes they define. We have 576 pairs in total, with class HH containing 444 pairs, class LH has 94 pairs, and class LL has 38 pairs. The mean value of RATE-PCC for each class is 0.156, 0.257, 0.322, respectively, and Figure 5.9a draws the distribution of RATE-PCC within the classes. ANOVA shows that the three classes do not differ significantly in

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



(a) RATE-PCC vs. RATE-NUM

(b) RATE-SIM vs. RATE-NUM

(c) NET-SIM vs. RATE-NUM

(d) NET-LHN vs. RATE-NUM

Figure 5.9: Classes based on RATE-NUM

Table 5.17: Mean RATE-PCC of RATE-NUM classes

		H	5	10	20	30	50	70	100
		HH	0.157	0.162	0.152	0.156	0.191	0.134	0.124
L	LL	LH							
5	0.93		0.68	0.66	0.743	0.705	0.613	0.573	0.49
10	0.322		0.233	0.236	0.273	0.257	0.302	0.355	0.254
20	0.19		0.201	0.208	0.217	0.204	0.206	0.24	0.192
30	0.2		0.19	0.19	0.191	0.192	0.21	0.216	0.182
50	0.19		0.172	0.176	0.17	0.174	0.18	0.191	0.164
70	0.176		0.167	0.171	0.17	0.17	0.182	0.198	0.164
100	0.185		0.17	0.172	0.172	0.174	0.185	0.193	0.164

terms of their mean RATE-PCC (p-value of 0.068). The conclusion is that classes based on RATE-NUM do not differ substantially in terms of their RATE-PCC.

Does RATE-SIM depend on RATE-NUM? We next consider whether there are

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

differences across RATE-NUM classes in terms of the RATE-SIM, instead of RATE-PCCs. Table 5.18 presents the mean RATE-SIM for different definition of classes.

Table 5.18: Mean RATE-SIM of RATE-NUM classes

		H	5	10	20	30	50	70	100
		HH	0.045	0.045	0.047	0.048	0.042	0.033	0.0077
L	LL	LH							
5	0.91		0.075	0.065	0.093	0.097	0.077	0.057	0.032
10	0.167		0.049	0.057	0.037	0.033	0.04	0.041	0.01
20	0.08		0.05	0.048	0.042	0.039	0.031	0.034	0.028
30	0.07		0.048	0.046	0.042	0.041	0.035	0.035	0.029
50	0.06		0.048	0.047	0.045	0.045	0.035	0.036	0.03
70	0.054		0.048	0.048	0.047	0.047	0.038	0.038	0.031
100	0.052		0.047	0.047	0.047	0.047	0.039	0.036	0.032

We fix L and H to their default values, and plot the distribution of RATE-SIM for each class in Figure 5.9b. In addition, we perform ANOVA and find that the means of classes differ significantly (p-value of $< 10^{-5}$). However, post hoc analysis, shown in Table 5.19, finds that the magnitude of the differences is not significant. Hence, we cannot draw any safe conclusions in this experiment.

Table 5.19: RATE-SIM differences across RATE-NUM classes

Pair	Diff. of Means	95% CI
LL - LH	0.1333	-0.0560,0.3226]
LL - HH	0.1195	[-0.0595,0.2984]
LH - HH	-0.0138	[-0.0515,0.0239]

Does NET-SIM depend on RATE-NUM? Next, we consider global network pairwise similarity between friends. Table 5.20 shows mean NET-SIM for the different definitions of RATE-NUM-based classes.

Table 5.20: Mean NET-SIM of RATE-NUM classes

		H	5	10	20	30	50	70	100
		HH	0.017	0.016	0.018	0.015	0.01	0.011	0.003
L	LL	LH							
5	0.35		0.02	0.02	0.021	0.022	0.023	0.02	0.0009
10	0.12		0.03	0.016	0.016	0.016	0.015	0.011	0.001
20	0.03		0.02	0.015	0.016	0.016	0.016	0.017	0.018
30	0.02		0.02	0.016	0.018	0.016	0.016	0.015	0.014
50	0.02		0.02	0.017	0.018	0.016	0.016	0.017	0.012
70	0.02		0.02	0.017	0.018	0.016	0.016	0.016	0.011
100	0.02		0.02	0.017	0.018	0.017	0.016	0.016	0.011

Again, we fix L and H to their defaults, and plot NET-SIM distributions for the three induced classes in Figure 5.8c. As before, while ANOVA shows that the means are

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

not equal with high significance (p-value of $< 10^{-13}$), post-hoc analysis, presented in Table 5.21, shows non-significant differences.

Table 5.21: NET-SIM differences across RATE-NUM classes

Pair	Diff. of Means	95% CI
LL - LH	0.1034	-0.0302,0.2369]
LL - HH	0.1047	[-0.0234,0.2328]
LH - HH	0.0013	[-0.0088,0.0114]

Does NET-LHN depend on RATE-NUM? The last experiment studies local network pairwise similarity between friends. Table 5.22 shows mean NET-LHN for the different definitions of RATE-NUM-based classes.

Table 5.22: Mean NET-LHN of RATE-NUM classes

		H	5	10	20	30	50	70	100
		HH	0.005	0.005	0.006	0.003	0.004	0.0008	0
L	LL	LH							
5	0.125		0.004	0.004	0.005	0.005	0.008	0.005	0
10	0.09		0.015	0.003	0.003	0.004	0.003	0.003	0
20	0.015		0.007	0.002	0.003	0.004	0.006	0.007	0.013
30	0.009		0.007	0.004	0.005	0.003	0.004	0.005	0.008
50	0.008		0.006	0.005	0.006	0.004	0.004	0.005	0.005
70	0.006		0.007	0.004	0.005	0.004	0.004	0.005	0.004
100	0.006		0.006	0.004	0.005	0.004	0.004	0.005	0.004

For fixed L and H, Figure 5.9d plots the distribution of NET-LHN in the three classes. ANOVA finds that they all have roughly equal means, and thus we conclude that no dependence on RATE-NUM is exhibited.

5.4 Conclusions

Based on the findings obtained from analysis for all pairs, RQ2 is answered, we see that rating similarity between pairs of users is related primarily with their network distance and with a particular metric of network similarity, SimRank. At both levels, rating and social behavior seem to be related, and specifically, rating seems to determine social behavior more strongly than the other way around.

Based on analysis for friends, RQ3 is answered we find out if a user with low social activity is connected with a user with high social activity, we expect their feedback similarity, in terms of RATE-PCC, to be almost two times as high as other pair of friends. Although we cannot be certain of the direction of influence, we conjecture that it flows from the more socially active user to the less active one.

The results obtained here could be exploited to provide more effective personalization. Specifically, we have found that to some extent network-based similarity can substitute

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

feedback-based similarity, and thus be used as a proxy for determining the similarity between friends in terms of their preferences. Moreover, the similarity strength increases when one friend is much less active than the other. These findings could be applied in a collaborative filtering approach, where tastes of similar minded users are aggregated. One idea would be to consider in this aggregation the strength of influence between two friends, computed based on their network similarity and their level of feedback activity.

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

CHAPTER 6

Analysis at Community Level

Community level (Level 3) assesses the behavior of users in a community. As a typical question is, do community members behave similarly in terms of rating behavior? or does being in the community affect users positively or negatively, in other words, do users enjoy communities or not? In pursuance of grouping users into communities we use two methods of community detection: Influencer based communities and Modularity based communities. To examine the community influence impact on users we implement several Recommender Systems with Social structure. We conduct this part of research using Douban dataset and we determine the usefulness and effectiveness of these models by using evaluation metrics such as novelty, diversity, Normalized Discounted Cumulative Gain (NDCG) and RMSE.

We begin with the research question that encompasses similarity of users in a community:

RQ4 How similar are the users in a community? How is overall similarity affected by different methods?

Similarity in terms of rating behavior (V1) quantifies how similar the ratings given by two users is. For this reason, we use the matrix factorization based similarity which is the inner product of users' latent factors. The next research question considers the notion of community influence (CI).

RQ5 How strong is the community influence? How is the community influence affected by different methods?

The community influence is defined and computed. Different statistical metrics are adopted to answer the RQ5, such as difference, relative difference and outliers detection.

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

In this chapter, two different mechanisms for community detection are used namely Influencer based communities (Inf-comm) and Modularity based communities (Mod-comm). The first mechanism creates a community around a user with a high degree centrality who we call influencer. The second mechanism is Mod-comm which is modularity based communities. In this case we use the greedy modularity communities algorithm to create communities.

6.1 Communities

6.1.1 Dataset

For this part of work we use Douban dataset [RSK16, MB15]. Douban is one of the large E-commerce platforms in China. Douban contains 129,490 unique users and 58,541 unique movie items. The total number of movie ratings is 16,830,839. For the social friend network, there are a total of 1,692,952 social relationships. Douban allows users to share their comments and viewpoints about movies. Users are able to post short or long comments on movies and give them marks. Two files are included in this Douban dataset, the user-item rating file (UserId, ItemId, Rating) and the user social friend network file (UserId1, UserId2). When this dataset was crawled Douban only allowed the Facebook-like friendship building mechanism but now Douban also supports the Twitter-like following mechanism.

6.1.2 Influencer based communities description

Inf-comm mechanism simply means that each community is formed by friends and friends of friends of an influencer. An influencer is one of the top 50 users with the highest degree. The size of the community indicates how many users are in the community. Descriptive analysis helps to determine the number of communities and the size of each community. In the end we get 48 communities after dropping two communities that were very small. Figure 6.1. shows the size by the community id. Figure 6.2 shows the number of influencers in each community. The visualization indicates that the bigger the size of the community the higher the chances of a community to have a high number of influencers. Among 48 communities, 6 communities have more than 120 users each. Community 4 has the highest number of user which is 198.

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

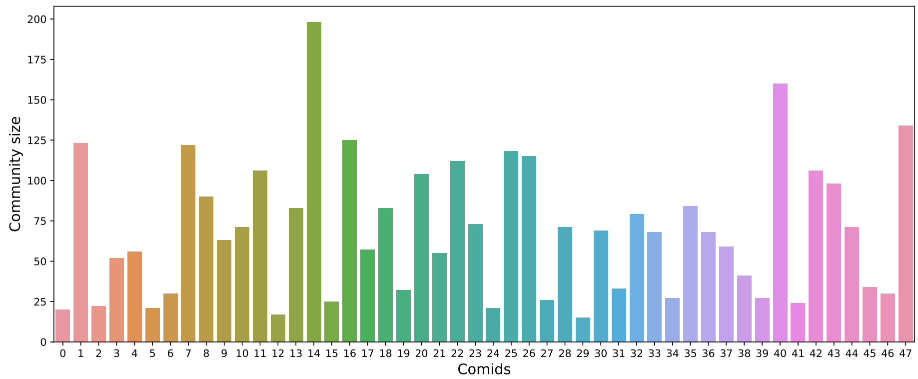


Figure 6.1: Number of members/community

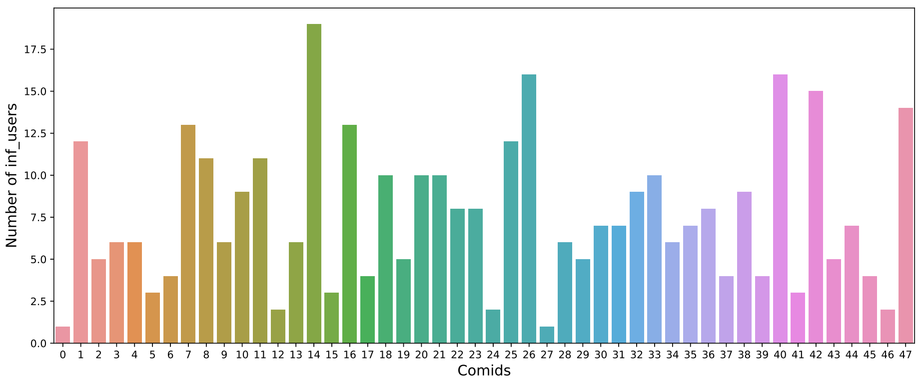


Figure 6.2: Number of influencers/community

6.1.3 Modularity based communities description

The community structure in the network is a well studied problem. The optimization by *Modularity* which is the quality function is one of the most effective mechanisms for possible partition in networks in general. The important key is it identifies groups embedded within a network by locating sets of nodes that interact with each other more frequently than the rest. By using this method we detect 5757 communities in total. For the purpose of doing the best analysis, only 31 communities are considered in our work because other communities have less than 10 community members. Figure 6.3 shows the size of the community. Influencer members are shown in Figure 6.4, where from community id 0 to the third community at least four members of of the community have the degree centrality value higher than ten, the same trend is observed in community 6.

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

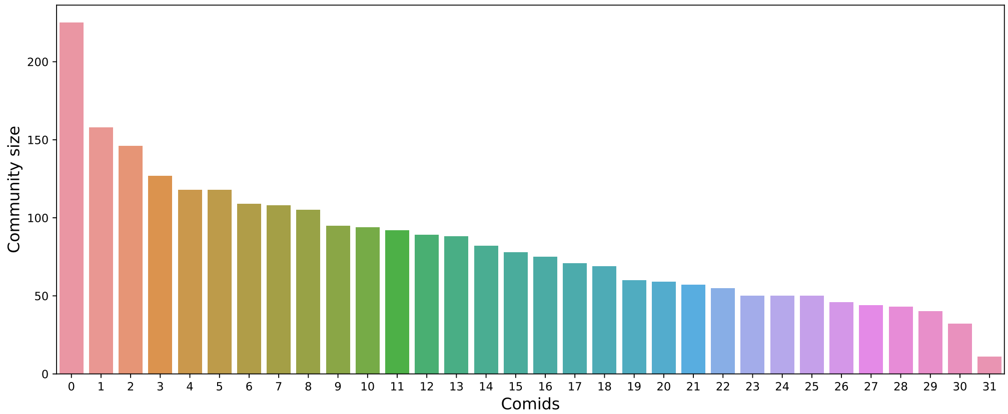


Figure 6.3: Number of members/community

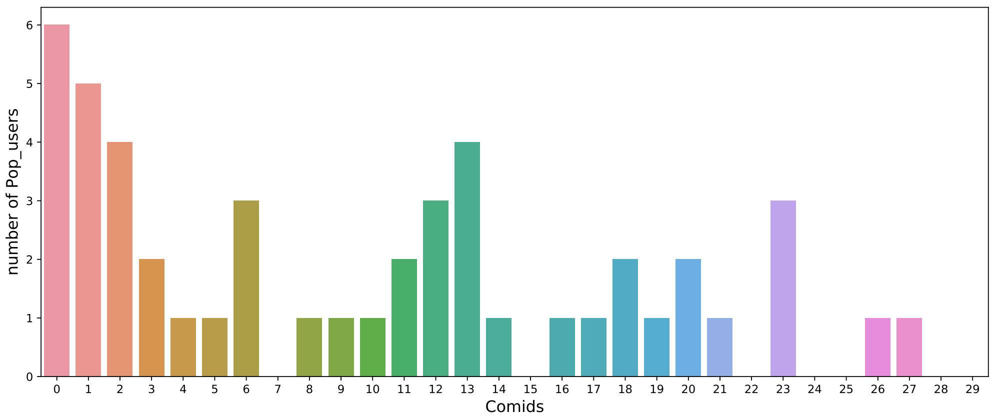


Figure 6.4: Number of influencer members/community

6.2 Methods

6.2.1 Models implementation

Matrices used by recommender

We adopt the idea of RS with with Social Regularization (RSR). RSR is defined as using social friends network to improve recommender systems, a way to model social network information as regularization terms to constrain the matrix factorization framework. In this model the training can be done in two different ways namely, *Average based*

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

Regularization and *Individual based Regularization* as described in Chapter 2. The following matrices are used for regularization:

- *S*: is an adjacency matrix S of the social graph.
- *Q*: is the rating similarity matrix (RATE-PCC); normalized to [0,1].
- *SQ*: is the adjacency matrix with rating similarities, element-wise product.
- *X*: is the NET-SIM (SimRank) node similarity matrix.
- *SX*: It is the adjacency matrix with Simrank (node similarity), element-wise product of S,X.
- *LH*: is the adjacency matrix of pairs of friends. Contains 1 when a pair is LH (low-high pairs) and 0 otherwise.
- *LH2*: is the adjacency matrix of pairs of friends. It considers how big is the difference of degree centrality between two users in a pair and score them from 0 to 1 similarity values.
- *HH*: is the adjacency matrix of pairs of friends. High-high pairs and represent a pair of inlfuencers.

Methods used

- *MF*: is the basic matrix factorization model used.

  Each model can be implemented on average and individual social regularization based. For example:

- *S*: this model, introduced in [JE10], extends MF by including average social regularization based on the adjacency matrix S of the social graph.
- *Si*: this model extends MF by including individual social regularization based on the adjacency matrix S of the social graph.

  Each method can be implemented by including one or more regularizers. For example:

- *SQ+LH2X*: the model includes SQ which results from an adjacency matrix with rating similarities, element-wise product. SQ is the first regularizer. The second regularizer is LH2X wich is element-wise product of the matrices LH2 and X.

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

Data Sampling

We only consider users that are present in both user-item rating (uir) file and social friendship (social) file. We performed 5-fold cross validation, splitting the dataset into train and test. For uir file we only select users with number of ratings greater than 2. As for social file we select users depending on their degree centrality and the requirements of the method we use in community formation and models implementation. The used sample consists of 1,048,575 ratings and 1,048,575 connections. The train set contains 838,710 and the test set contains 207,835, for the social we have 7,908 connections.

Setting and Utilization of Hyperparameters

The batch size (bs) is a hyperparameter that defines the number of samples to work through before updating the internal model parameters. This is done through changing from one sample to another and making predictions. At the end of the batch, an error is calculated by comparing the predictions with the output variables. The algorithm is then updated basing on this error so as to improve on the model.

A training dataset can be divided into one or more batches.

- *Batch Gradient Descent*, Batch Size = Size of Training Set
- *Stochastic Gradient Descent* , Batch Size = 1
- *Mini-Batch Gradient Descent* , 1 < Batch Size < Size of Training Set.

The size of a batch must be more than or equal to one and less than or equal to the number of samples in the training dataset. We set the bs to 32,768 which is 2^{15}

The number of epochs (ne) is a hyperparameter that defines the number of times that the learning algorithm will work through the entire training dataset. A learning algorithm is said to have one epoch when each sample in the training dataset has had an opportunity to update the internal model parameters. An epoch is comprised of one or more batches. For example, as above, an epoch that has one batch is called the batch gradient descent learning algorithm. The ne is set to 20.

Number of factors (nf): In matrix factorization concept the idea is to factorize a matrix, i.e. to find out two (or more) matrices of lower dimensionality such that when by multiplying them you can get back the original matrix. The objective is to discover latent features/factors. Increasing the number of latent factor will improve personalization, therefore recommendation quality, until the number of factors becomes too high, at which point the model starts to overfit and the recommendation quality will decrease. Number of factor is the dimensionality. The nf we use are 10, 20, 50, 100 and 200

Weight decay (wd) is an additional term in the weight update rule that causes the weights to exponentially decay to zero, if no other update is scheduled. Weight decay

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

penalizes model complexity, so it is used to control model's variance against its bias. Clearly if the complexity is penalized too much, the model will not learn anything useful, since it will be too simple. The ranges tested are 5e-5, 2e-5, 1e-5, 5e-6, 2e-6 and 1e-6.

The learning rate (lr) determines how much an updating step influences the current value of the weights. A small learning rate may be "under-fitting"(or the model has "high bias "), and a large learning rate may be "over-fitting"(or the model has "high variance"). The evidence of under-fitting is both training and testing set have large error, and the error curve for training and testing are close to each other. The sign of over-fitting is training set's error is very low and testing set is very high, two curves are far away from each other. The ranges tested are 5e-1, 2e-1, 1e-1, 5e-2, 2e-2 and 1e-2.

6.2.2 Models selection

Here we discuss how we select hyperparameters values. We only show the case of MF. Weight decay (wd) and learning rate (lr), vary depending on nf. Table 6.1 and Figure 6.5 below describe how wd and lr were determined.

Table 6.1: Best values for basic MF

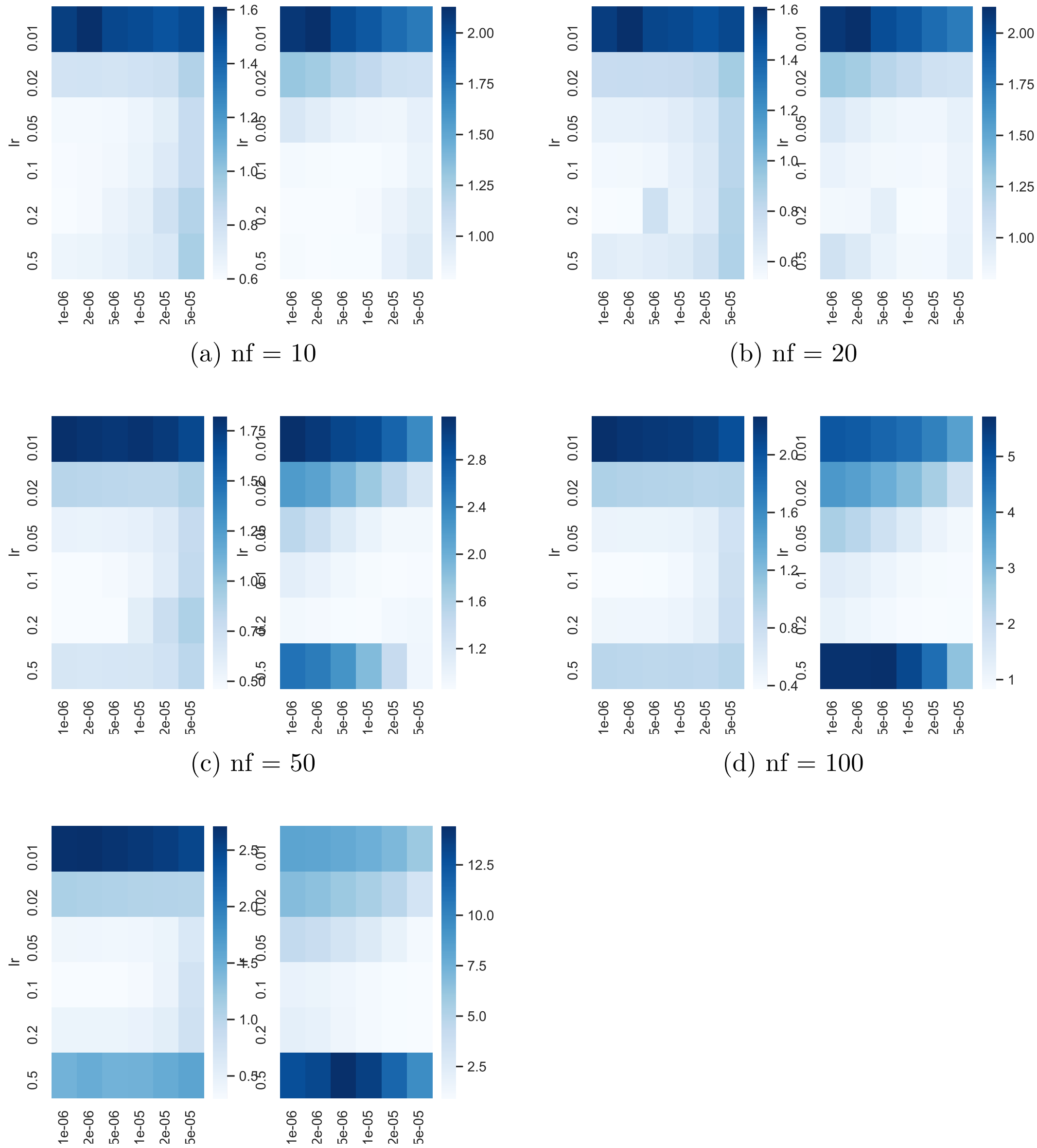
	nf = 10	nf = 20	nf = 50	nf = 100	nf = 200
lr	2e-1	2e-1	2e-1	2e-1	2e-1
wd	5e-6	1e-5	5e-6	2e-5	5e-5

Test/train split

In order to set hyperparameters, a range of values for learning (lr) rate and weight decay (wd) is first determined. It is observed that there is an impact on rmse training set and rmse test set for each value that is tried within the range. To choose the best parameter values we pick the values that maintained the lower RMSE in test. Figure 6.5 shows the best values.

The paired heat maps below present our selection process of the best values of the weight decay (wd) and learning rate (lr) for each n-factor (nf). The paired heat map as shown in Figure 6.5 represents the RMSE for a given range of wd and lr: the left heat map of the pair represents the RMSE computed on the train data, while the right heat map of the pair represents the RMSE computed on the test data. The dark blue color indicates the highest values and the light blue color stands for the lowest values. In this particular case, we are interested in the lowest RMSE values which we can find in both the left and the right heat maps. For example, in Figure 6.5a. where nf = 10, we have the best values (lr = 2e-1 and wd = 5e-6). We can see that those are the lightest grid for both train and test data.

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



(a) nf = 10
(b) nf = 20
(c) nf = 50
(d) nf = 100
(e) nf = 200

Figure 6.5: Hyperparameters (wd,lr) best values for every nf (DOUBAN)

6.2.3 Models results interpretation

In order to evaluate the performance in terms of accuracy ,effectiveness and usefulness of the recommenders the following metrics are applied:

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

RMSE

RMSE is computed for every model at different nf, nf here is a value representing n-factors that is set in the range of 10, 20, 50, 100, 200. Figure 6.6 shows RMSE comparison chart. Table 6.2 shows the ranking rmse the sorted order. The top 15 models include our proposed HH (a pair of popular users) where influencers are taken into account, LHX where low-high (an unpopular and a popular pair) pairs and SimRank similarity values are considered, and also LH2X.

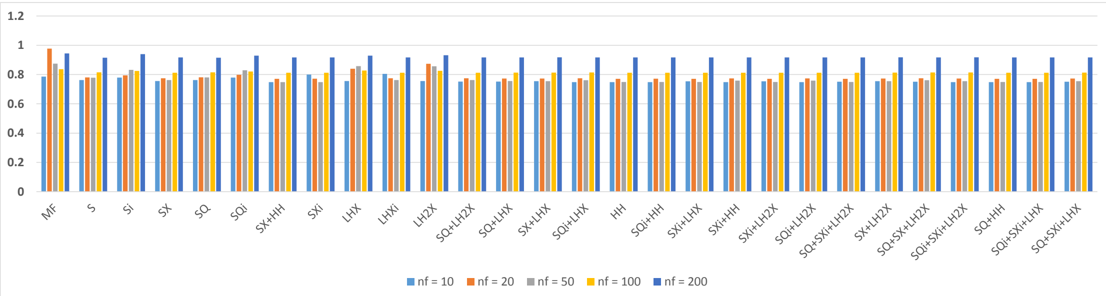


Figure 6.6: RMSE comparison chart of models per nf range

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

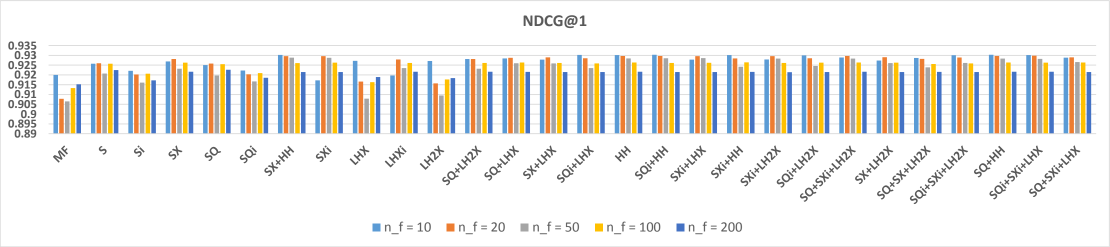
Table 6.2: Ranking Models by RMSE

Models	**df = 10**	**df = 20**	**df = 50**	**df = 100**	**df = 200**
MF	0.7864	0.9771	0.8747	0.8368	0.9444
S	0.7619	0.7802	0.7783	0.8151	**0.9147**
Si	0.7794	0.7944	0.8318	0.8249	0.9394
SX	0.7559	0.7745	0.7620	0.8124	0.9168
SQ	0.7624	0.7810	0.7806	0.8152	**0.9147**
SQi	0.7797	0.7981	0.8297	0.8211	0.9289
SX+HH	**0.7481**	**0.7712**	**0.7483**	0.8125	0.9172
SXi	0.8000	0.7715	0.7484	0.8124	0.9171
LHX	0.7559	0.8397	0.8573	0.8272	0.9287
LHXi	0.8045	0.7748	0.7623	0.8124	0.9168
LH2X	0.7559	0.8732	0.8570	0.8257	0.9321
SQ+LH2X	0.7522	0.7742	0.7622	0.8125	0.9168
SQ+LHX	0.7520	0.7728	0.7556	0.8132	0.9173
SX+LHX	0.7546	0.7728	0.7553	0.8132	0.9174
SQi+LHX	0.7484	0.7743	0.7611	0.8139	0.9172
HH	0.7486	0.7713	0.7490	0.8123	0.9170
SQi+HH	0.7482	0.7715	0.7489	0.8125	0.9171
SXi+LHX	0.7538	**0.7712**	0.7495	**0.8123**	0.9170
SXi+HH	0.7482	0.7738	0.7586	0.8125	0.9168
SXi+LH2X	0.7541	0.7713	0.7496	0.8124	0.9171
SQi+LH2X	0.7485	0.7733	0.7584	0.8124	0.9169
SQ+SXi+LH2X	0.7513	0.7713	0.7491	0.8124	0.9171
SX+LH2X	0.7547	0.7726	0.7545	0.8134	0.9174
SQ+SX+LH2X	0.7510	0.7749	0.7618	0.8141	0.9173
SQi+SXi+LH2X	0.7485	0.7726	0.7550	0.8140	0.9175
SQ+HH	0.7482	**0.7712**	0.7493	0.8125	0.9171
SQi+SXi+LHX	0.7486	0.7713	0.7491	**0.8123**	0.9171
SQ+SXi+LHX	0.7513	0.7726	0.7548	0.8133	0.9173

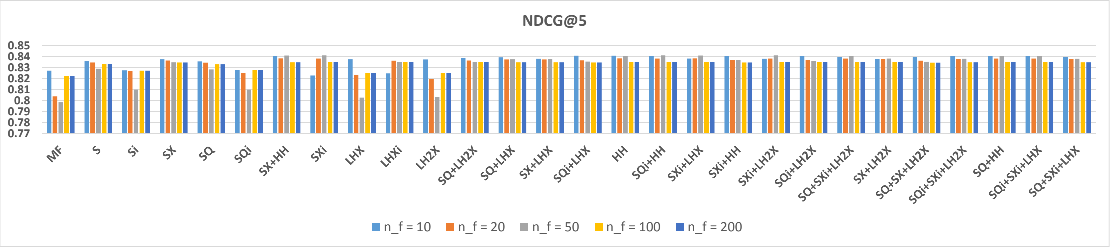
NDCG

The Normalized Discounted Cumulative Gain in order to measure usefulness, effectiveness of search algorithms and the ranking quality from range 1 to 20. Figure 6.7 shows the NDCG@k within the range.

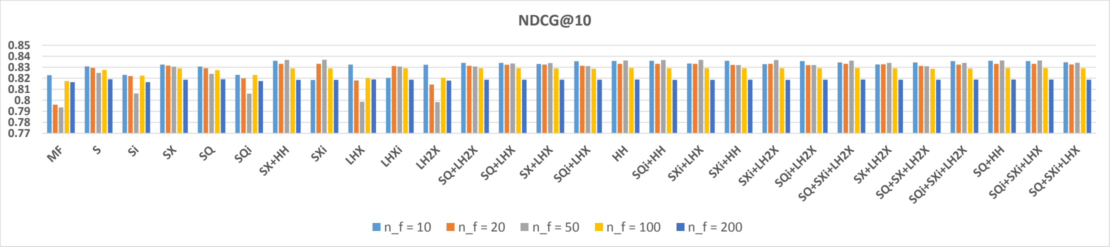
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub



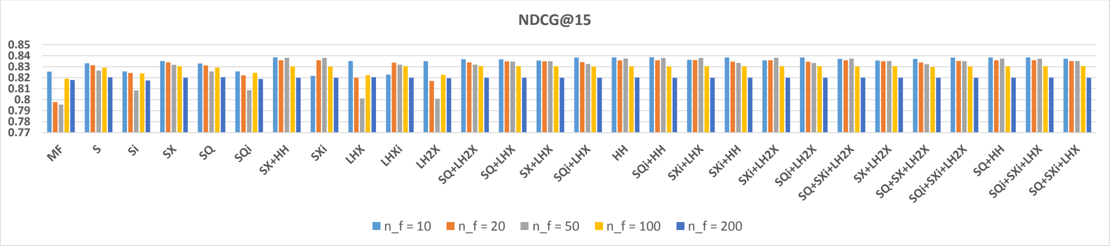
(a) NDCG@1



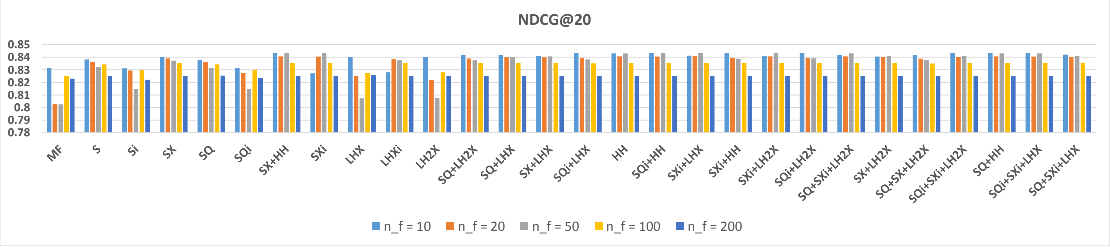
(b) NDCG@5



(c) NDCG@10



(d) NDCG@15



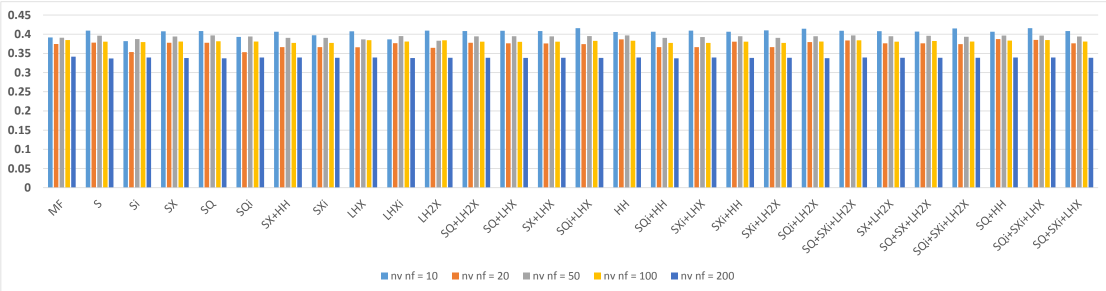
(e) NDCG@20

Figure 6.7: NDCG@k

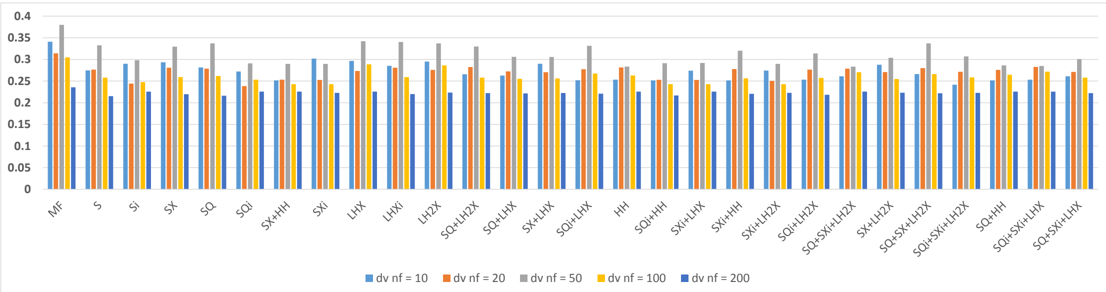
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Diversity and Novelty

In the past, many researchers came to realize that accuracy as evaluation metric alone is not enough to identify the effectiveness of a recommendation functionalities. Diversity and novelty metrics are the most important key qualities beyond accuracy in real recommendation scenarios when it comes to measure the utility.



(a) user_Novelty chart for models per nf range



(b) user_Diversity chart for models per nf range

The relation between both is that the diversity specifies how the set of items is diverse which means how different those items are to each other. Novelty means how different are the items with respect to past rated item. Figure 6.8b displays the results of diversity for nf range 10, 20, 50, 100 and 200. and Figure 6.8a shows the results of novelty in the same range of nf. Based on Novelty and diversity relationship as mentioned above. Figure 6.8a shows that our models outperform the existing models in terms of user-Novelty especially the base MF model. User-Diversity results (Figure 6.8b) shows that items presented to the users are more diverse with respect to each other for the recommenders with social structure compared to the basic MF recommender. The novelty indicates how different a piece of information is with respect to "what has been previously seen".

6.3 Definitions

6.3.1 Matrix Factorization based similarity

For the rating view, the MF-SIM is used to quantify user's rating similarity. MF-SIM is computed by using user factors (lantent factors). The MF-SIM computes the similarity

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

for every user to all others by computing inner products of user's lantent factors. The MF-SIM of a user U_i and a user U_f:

$$Sim(U_i, U_f) = \frac{U_i.U_f}{||U_i||||U_f||}$$

6.3.2 Community Influence (CI)

It reveals a percentage of the top-k most similar users to me within my community. CI (which we can also think of as the precision) is computed by using the following steps:

- *Ground truth* is a ranked list of the top most similar users to a selected user at a certain k, with k in range of 10, 20, and 50.
- *Community Id* is located: the community of a selected user
- *Create Hits* where Hit value (Phit) = 1, if a user index/id in the ranked list at k of the top most similar users to a selected user appears in the community id's list of members and Hit value (Nhit) = 0, otherwise

A parameter k is used to check and observe different levels of influence $CI_X@k$. This is done by either increasing or decreasing parameter k. The $CI_X@k$ is given by the formula below :

$$CI_X@k(u, c) = \frac{tp}{|c|}$$

where true positives tp the total number of users from my top k similar list who also belong to my community and $|c|$ is the community size.

Delta (Δ) of the Community Influence

Delta $\Delta CI_X@k(u; c)$ stands for the difference of after a community integration minus before the community integration.Delta indicates at what degree the CI of each community member changes is affected by the community compare to the baseline (MF: before the community integration). In a Community based on the method used Δ of a particular member of the community using specific method X at a certain k value (@k) is defined as follows:

$$\Delta CI_X@k(u; c) = CI_X@k(u; c) - CI_{MF}@k(u; c)$$

while $\Delta CI_X@k(c)$ of a community using specific method X at a certain k value (@k) is defined as follows:

$$\Delta CI_X@k(c) = CI_X@k(c) - CI_{MF}@k(c)$$

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

Relative Difference of the Community Influence

$r\Delta CI_X@k(c)$ of a community using specific method X at a certain k value (@k) is defined as follows

$$r\Delta CI_X@k(c) = \frac{CI_X@k(c) - CI_{MF}@k(c)}{CI_{MF}@k(c)}$$

$$= \frac{\Delta CI_X@k(c)}{CI_{MF}@k(c)}$$

Outliers of Delta

Outliers figuratively lay outside the rest of the data, they can be determined as positive outliers if they are in the range higher than the upper end or negative outliers if they are located in the range below the low end. Outliers can indicate an extremely positive or negative influence in our case. To compute an outlier the first step is to measure the spread which is given by the InterQuartile Range or $IQ_{\Delta CI_X@k(u;c)}R$ (range the between the quartiles).

$$IQ_{\Delta CI_X@k(u;c)}R = Q3_{\Delta CI_X@k(u;c)} - Q1_{\Delta CI_X@k(u;c)}$$

The size of the IQR indicates how spread out the the middle half of the data is. IQR reveals how far a typical value could be from the mean, which means anything much more than the typical distance can be spotted. We use the $1.5 \times IQ_{\Delta CI_X@k(u;c)}R$ rule to define outliers where:

Negative outiers:

$$= Q1_{\Delta CI_X@k(u;c)} - 1.5IQ_{\Delta CI_X@k(u;c)}R$$

Positive outiers:

$$= Q3_{\Delta CI_X@k(u;c)} + 1.5IQ_{\Delta CI_X@k(u;c)}R$$

6.4 Results

6.4.1 Distribution of Matrix factorization based similarity (MF-SIM)

In this analysis we use four different methods of RS with social regularization, namely S, SQ, LHX and MF, which is the baseline. The S method represents an adjacency matrix of the social network. The SQ method is a product of a social matrix S and a similarity matrix Q. The LHX gives a SimRank value for each pair of a popular and unpopular user. Matrix factorization based similarity (MF-SIM) is computed for every pair of users in a community for each method. Figure 6.9 shows the distribution of MF-SIM for all pairs of users. Furthermore, Figure 6.10 shows the distribution of MF-SIM for pairs of friends in social network. These distributions are computed for all the methods mentioned.

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

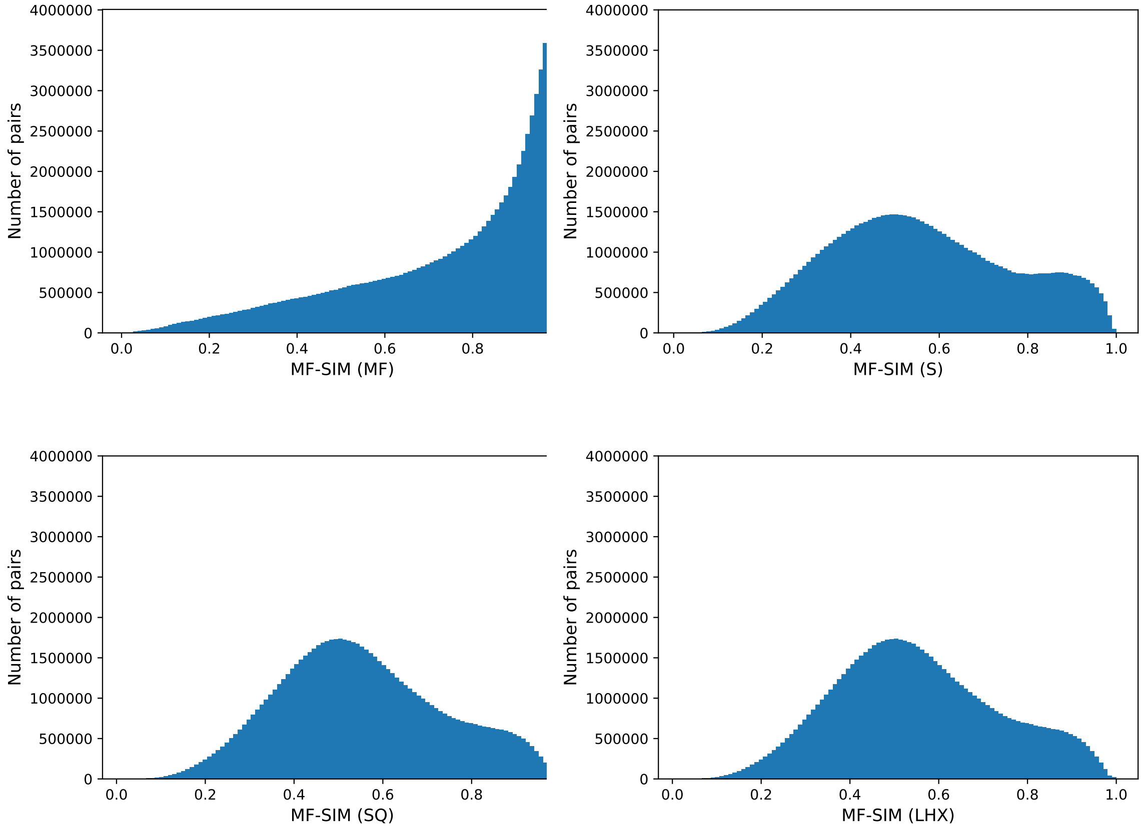


Figure 6.9: MF-SIM distribution for all user pairs

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

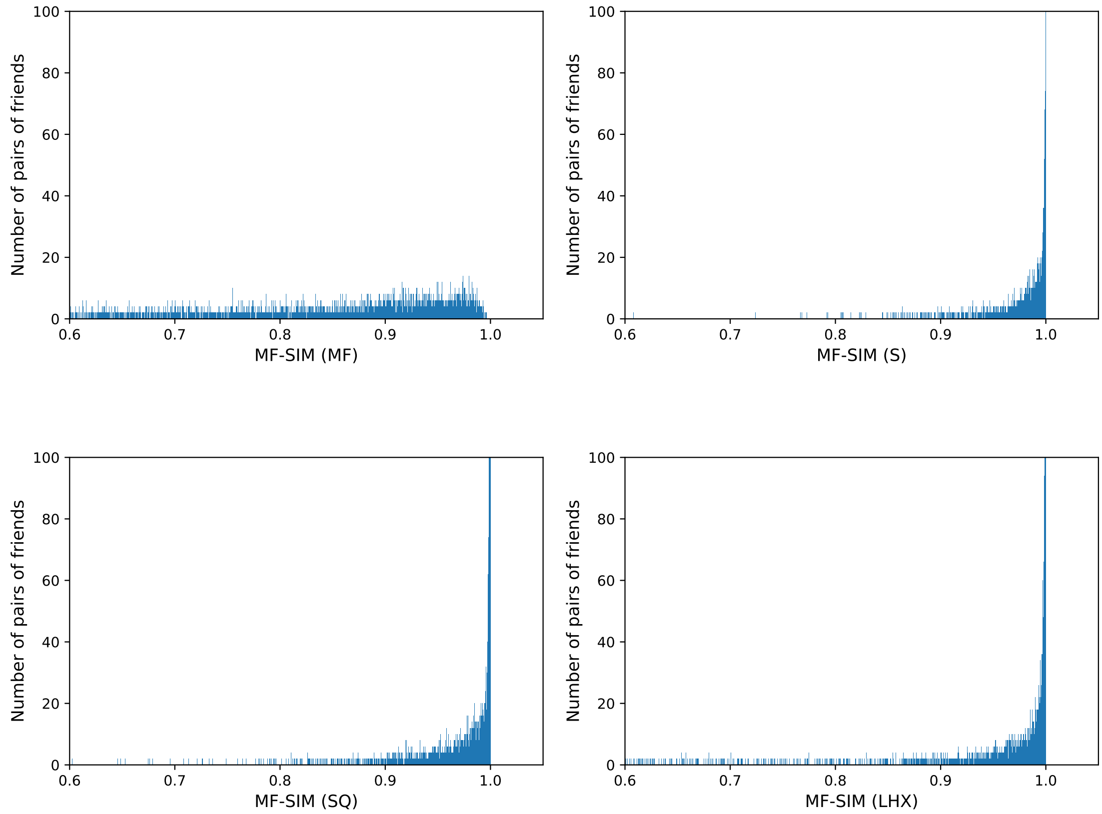


Figure 6.10: MF-SIM distribution for pairs of friends)

MF-SIM for Pairs of Friends in Communities

MF-SIM average

For further analysis, we compute the mean of the MF-SIM distribution in order to measure its central tendency for pairs of friends by both of the community methods, Influencer based communities and Modularity based communities. Figure 6.11 and Figure 6.12 plot the average of the MF-SIM based on all the methods (S, SQ, LHX, and MF). For every community, we observe that the mean of the MF-SIM is high, especially for recommenders with a social structure (S, SQ, and LHX). For them the mean value is higher than 0.90 which implies a very high MF-SIM value for the users, while for the MF method it is not the case. By way of conclusion, we summarize our main observation: similar users tend to be socially close, i.e., tend to form communities.

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

Figure 6.11: MF-SIM mean for pairs of friends for Inf-comm

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

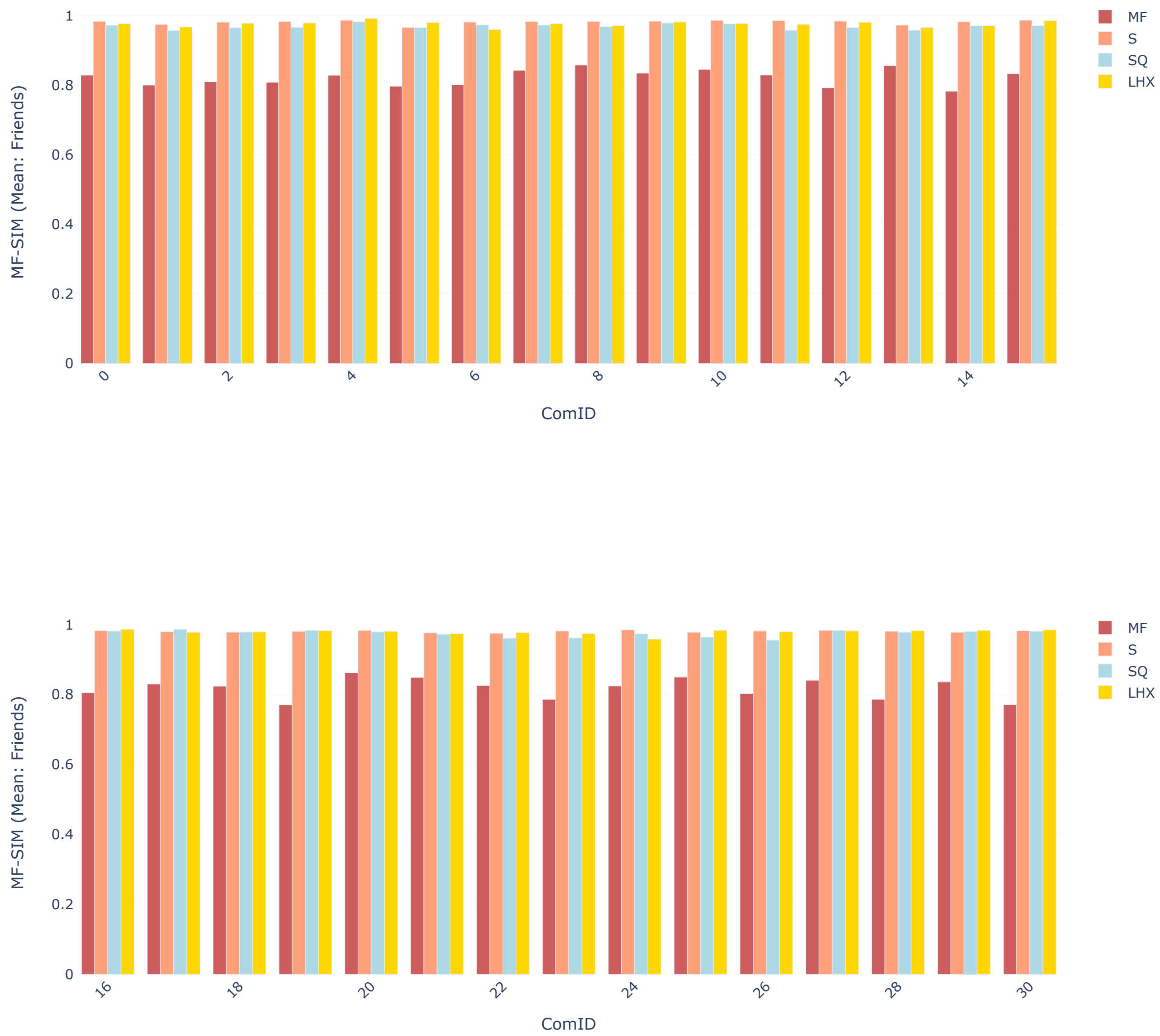


Figure 6.12: MF-SIM mean for pairs of friends for Mod-comm

MF-SIM variance

In addition to the MF-SIM mean, we also compute the MF-SIM variance by taking into account only pairs of friends inside communities. High variance indicates that numbers in a set are far from the mean and from each other. We observe that phenomena in Figure 6.13 (Influencer based communities) and Figure 6.14 (Modularity based communities) where

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

the MF-SIM variance is higher for the MF method than for other methods where the social structure is considered (S, SQ, LHX).

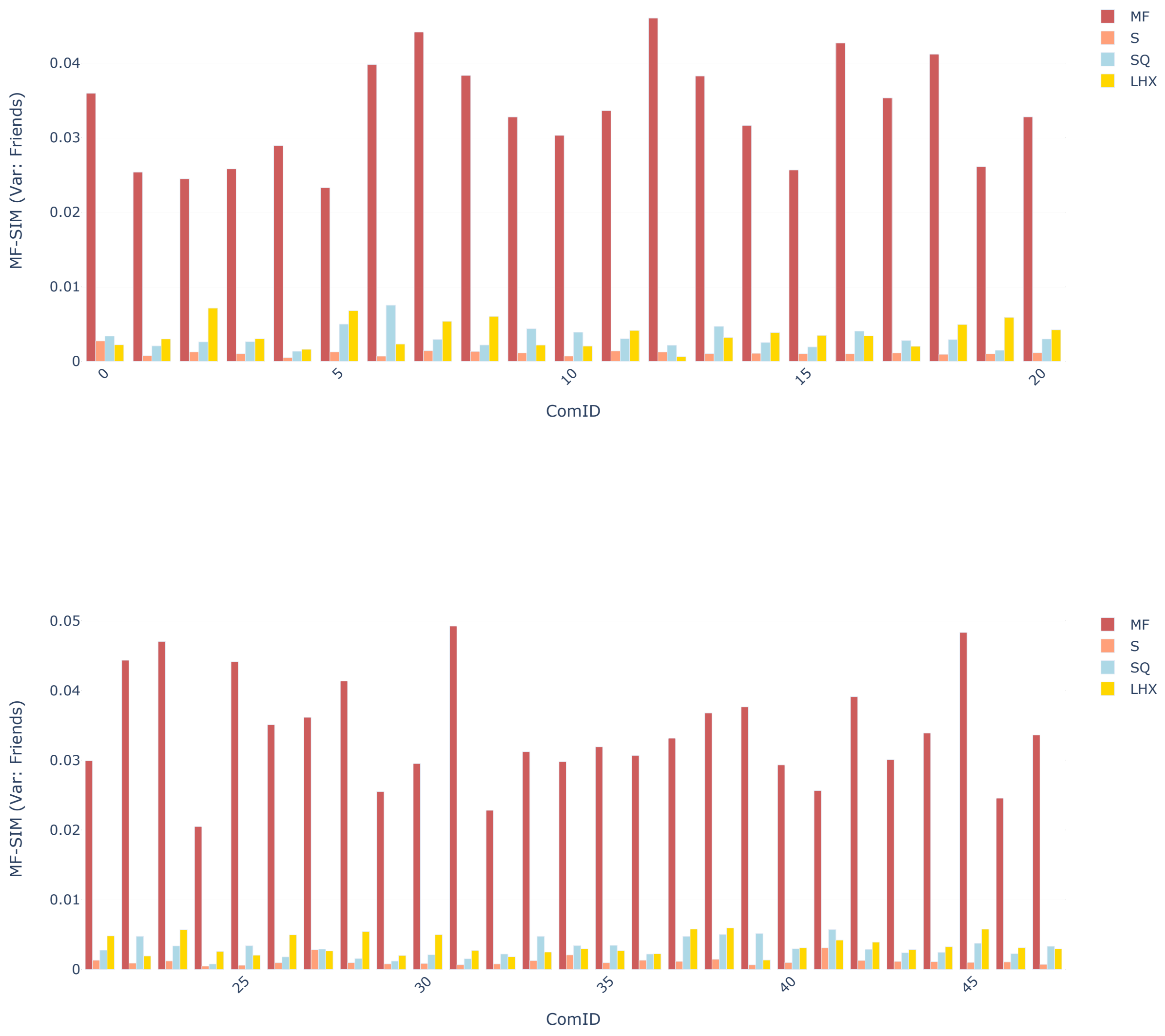


Figure 6.13: MF-SIM variance for pairs of friends for Inf-comm

This shows that there is a remarkable dissimilarity among users for the MF method. A very small variance is observed for the S and SQ methods which indicates that users in this recommenders are very similar within their communities.

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

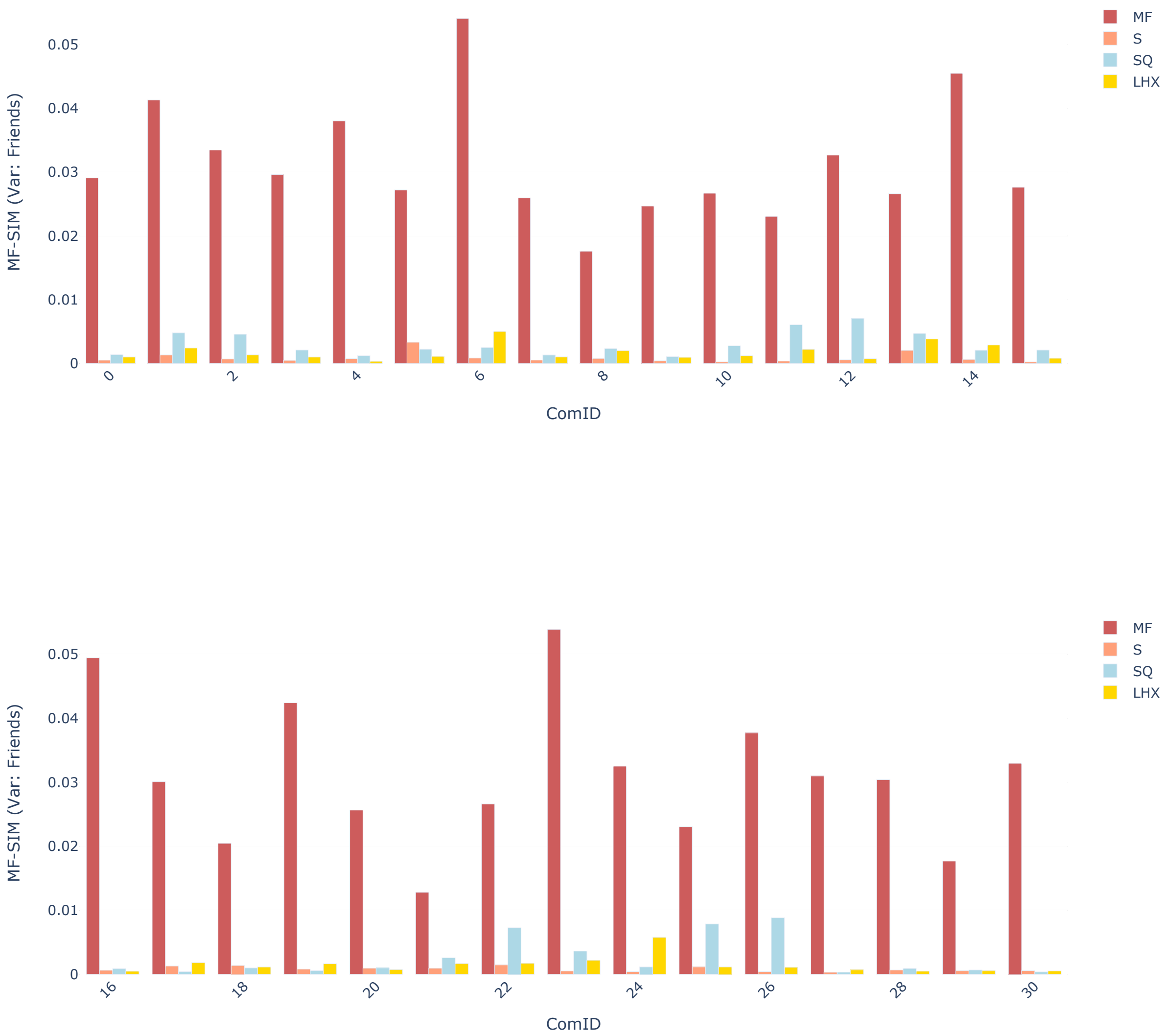


Figure 6.14: MF-SIM variance for pairs of friends for Mod-comm

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

MF-SIM distribution

In order to capture more information about the MF-SIM distribution and to study its characteristics in both of the community methods, we make use of boxplots. Boxplots show overall patterns of the MF-SIM for community members based on a specific recommender method.

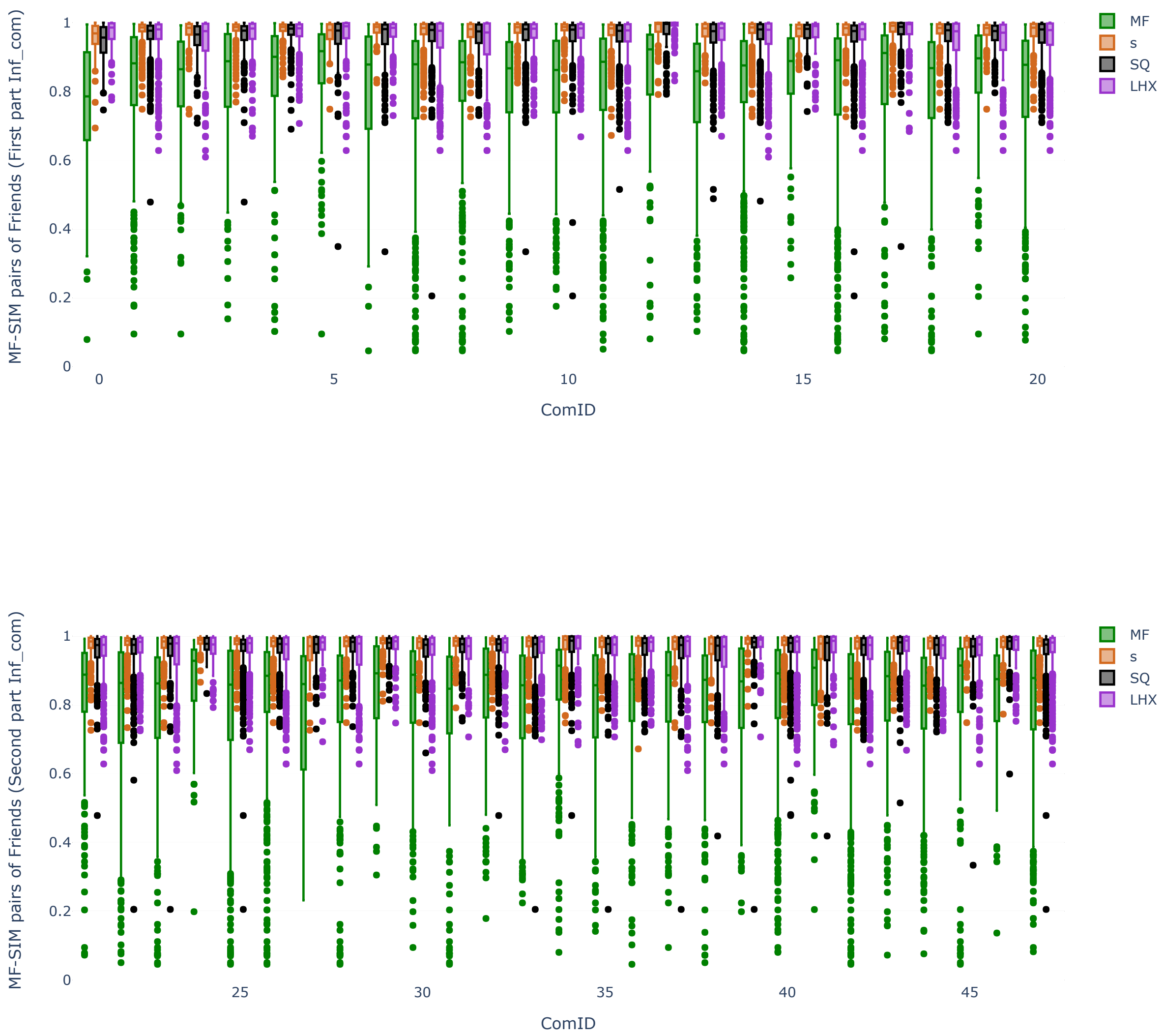


Figure 6.15: MF-SIM distribution for pairs of friends for Inf-comm

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

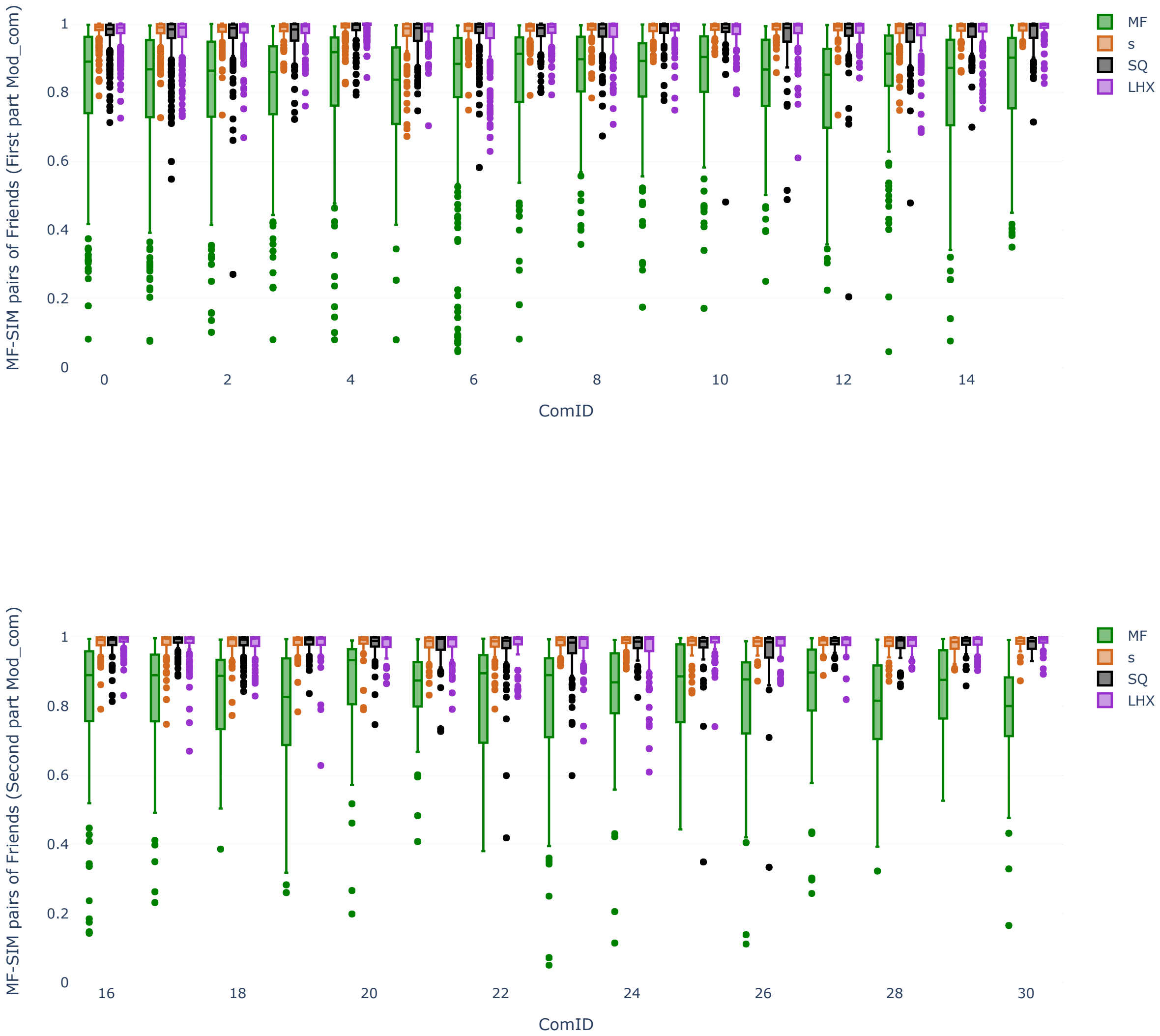


Figure 6.16: MF-SIM distribution for pairs of friends for Mod-comm

To interpret the range and other characteristics of the MF-SIM for a large group of community members based on different recommender methods, namely MF, S, SQ, LHX, we make the following conclusions:

Figure 6.15 depicts the results of the S, SQ, LHX methods for Influencer based communities. They suggest that overall community members have a high level of similarity

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

among each other. Therefore, when a recommender incorporates a social structure, community users are forced to become more similar. Focusing on the LHX method, we notice that among social recommenders the LHX has the least impact. This applies for both Influencer based communities and Modularity based communities.

Figure 6.16 illustrates the outcome of the MF method for Modularity based communities. The MF method (in green) shows various levels of similarity within a community. This leads to the conclusion that community members hold quite different MF-SIM values which means users are not similar. By taking into account the social structure to build a recommender, we trigger a high similarity among community members, especially when these members are friends, i.e., have strong social connections.

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

6.4.2 Community Influence for Individual View

Community Influence at k (CI@k)

What does CI impart?

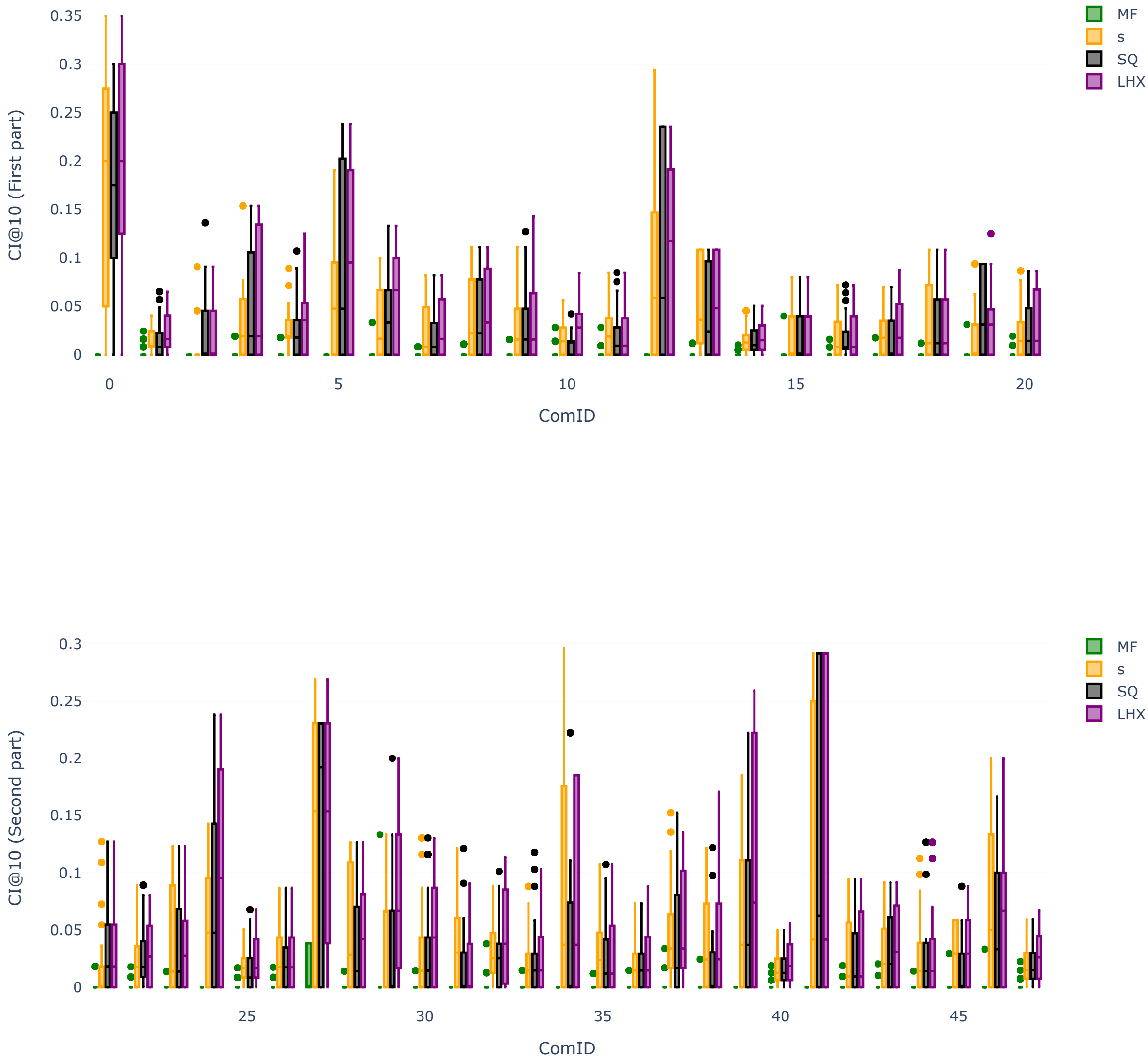


Figure 6.17: CI@10 for Inf-comm

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

Here we explore the influence within each community. For example, my community influence can be simply interpreted as a percentage of the top-k most similar users to me within my community.

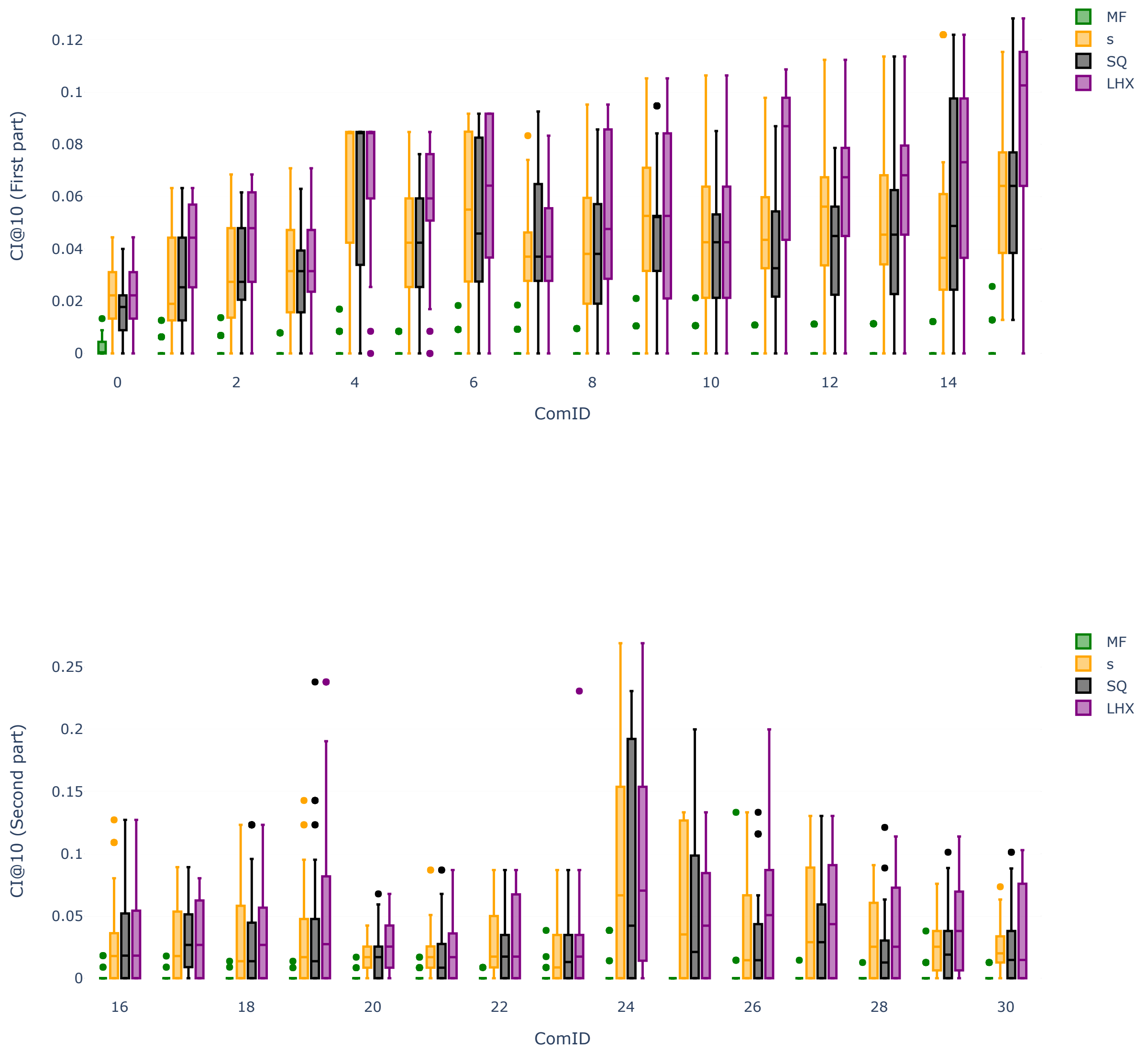


Figure 6.18: CI@10 for Mod-comm

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

When computing CI@k which is considered as a precision at k, we look at the number of predicted members of my community intersected with the ground truth (top-ranked list of the MF-SIM for the most similar users to me).

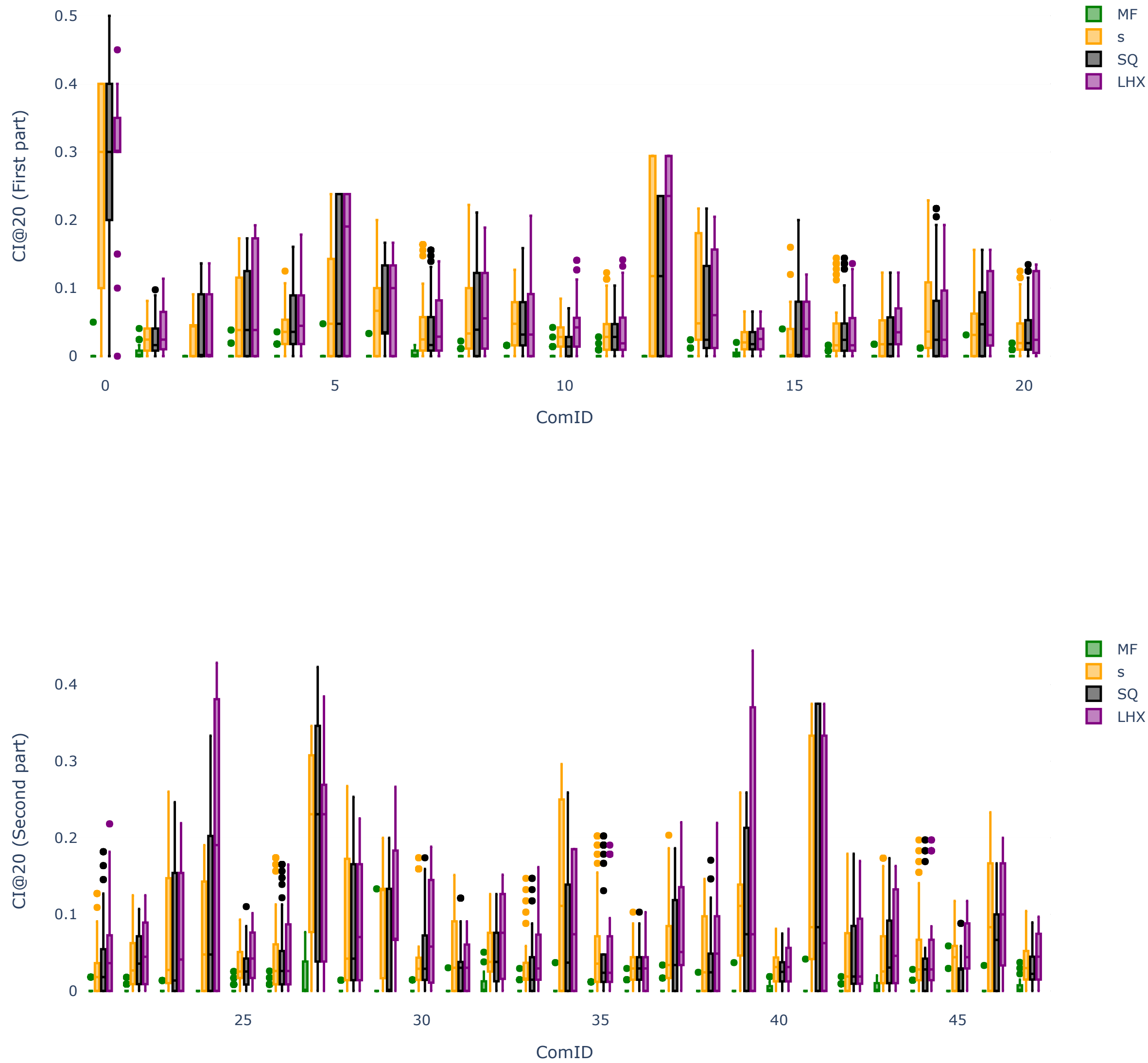
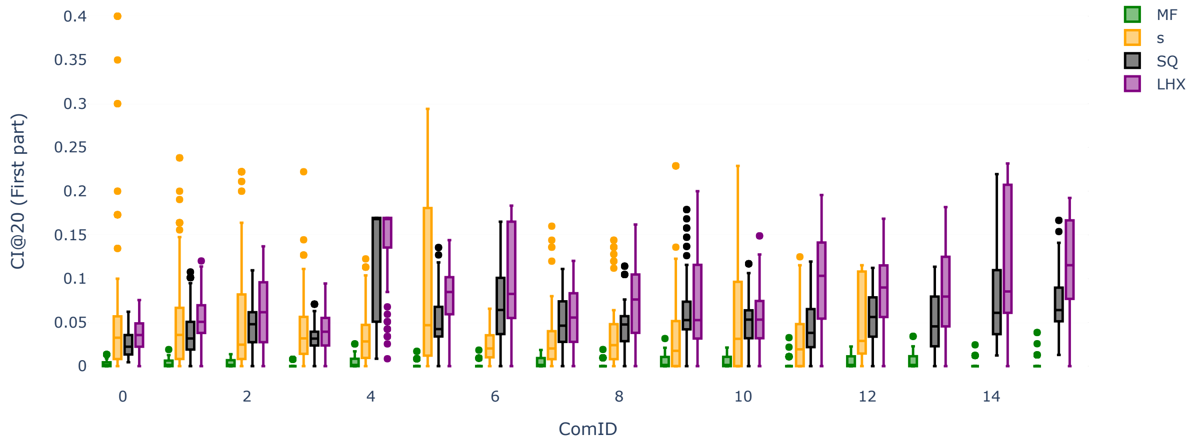


Figure 6.19: CI@20 for Inf-comm

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

CI is calculated for every community method (Influencer based community and Modularity based community) and for every k value (10, 20 and 50). Figure 6.17 for Influencer based community and Figure 6.18 for Modularity based communities compare the boxplots of the CI of the S, SQ and LHX methods with the MF method baseline for CI@10. It is noticeable that CI increases for recommenders that consider social structure.



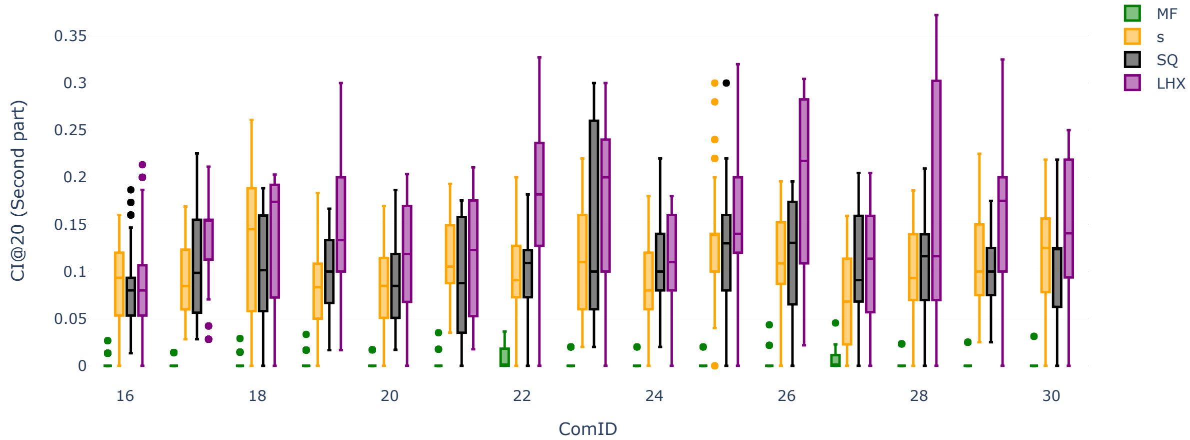
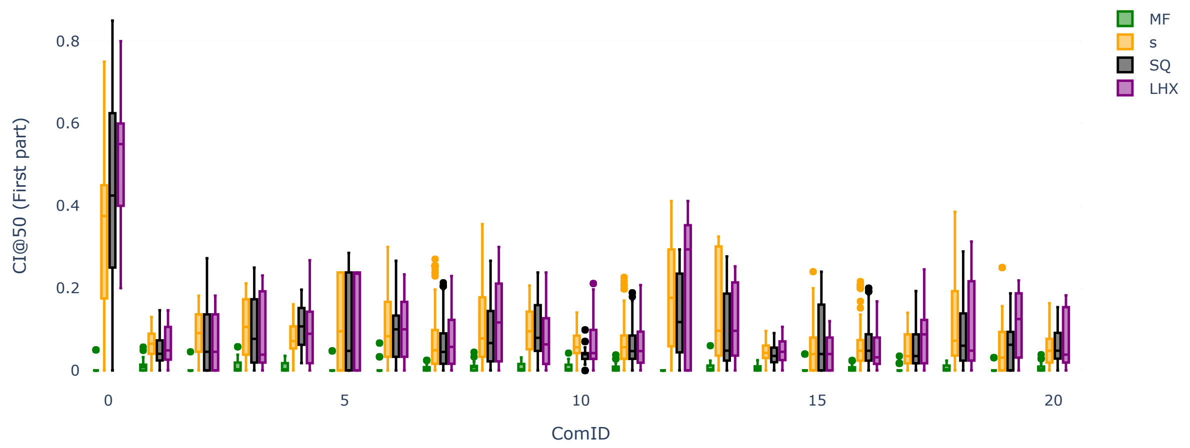


Figure 6.20: CI@20 for Mod-comm

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



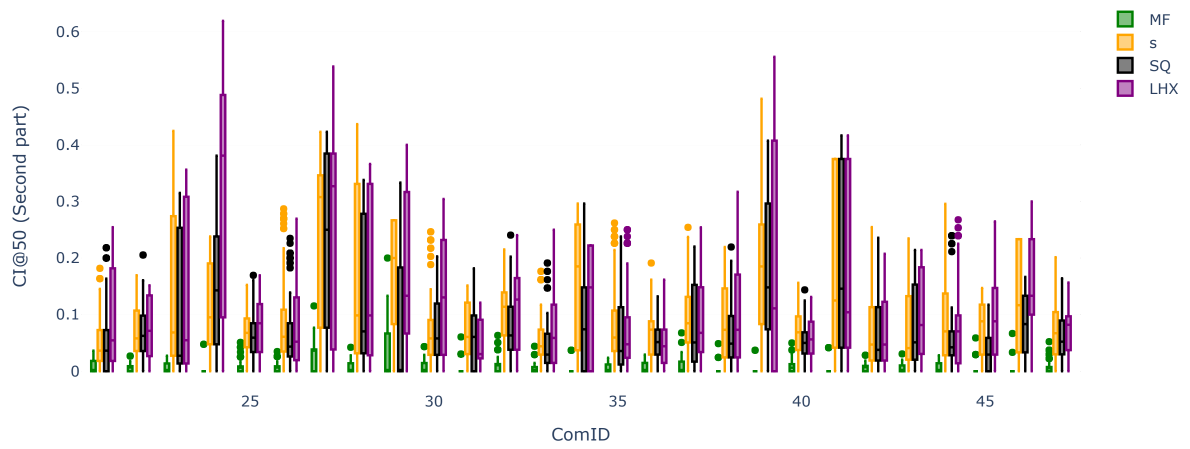


Figure 6.21: CI@50 for Inf-comm

In particular for the LHX method, the increase of CI is clearly observed. Figure 6.19 for Influencer based community and Figure 6.20 for Modularity based communities compare the boxplots of the CI of the S, SQ and LHX methods with the MF method for CI@20. The same trend is noticed for both community methods and recommender methods.

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

For the last figures (Figure 6.21 for Influencer based community and Figure 6.22 for Modularity based communities for CI@50), the significant increase of CI among the recommender methods with social structure and the MF baseline method with no social structure is observed as well. The LHX method has a significant growth of CI at every k.

Figure 6.22: CI@50 for Mod-comm

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

Delta (Δ)

What does DELTA (Δ) CI impart?

Delta indicates at what degree a CI of each community member changes compared to the MF baseline. The main goal of Delta is to detect users' behavior change before joining communities (or being integrated into social network) and after joining communities.

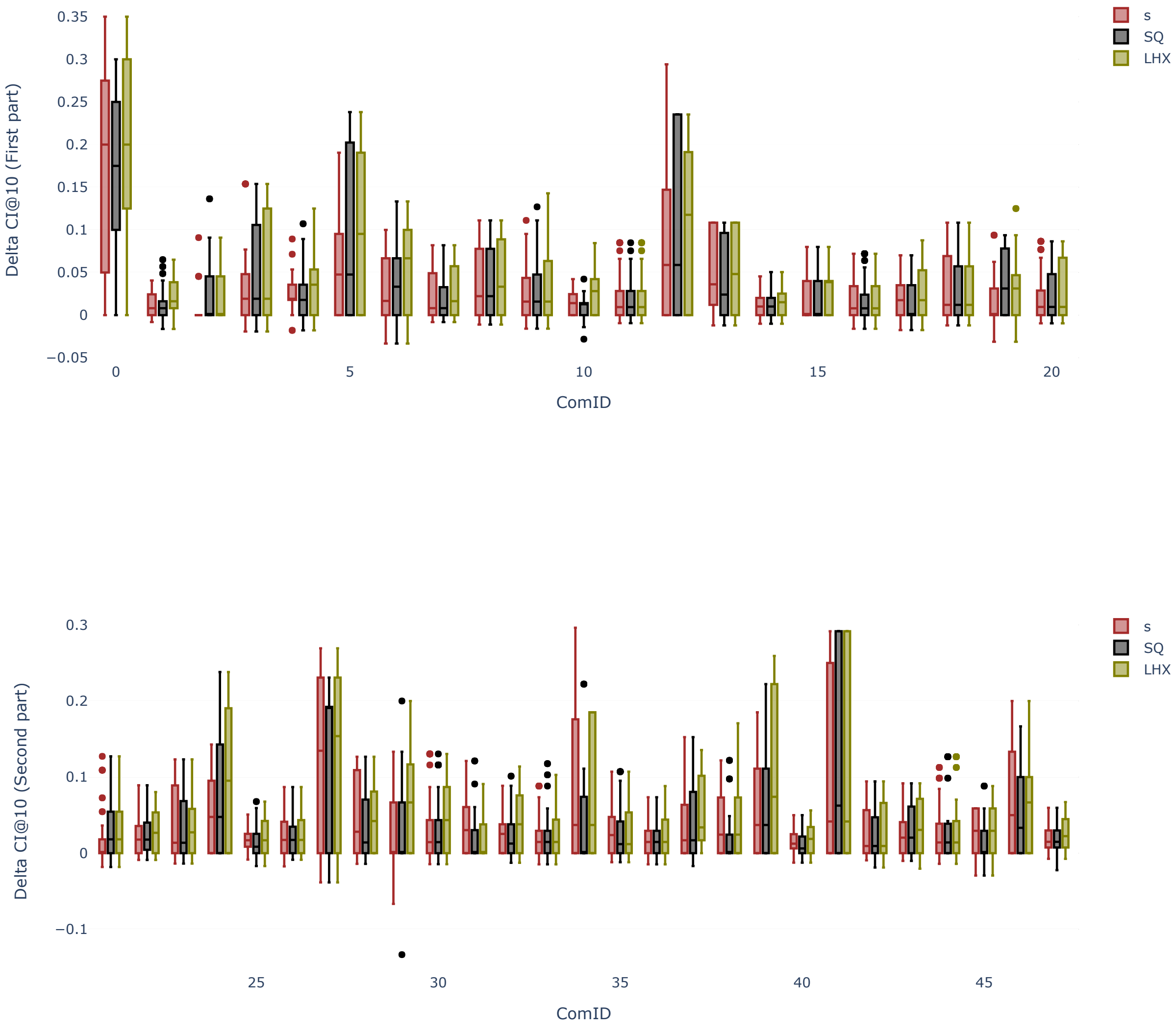


Figure 6.23: Delta CI@10 for Inf-comm

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

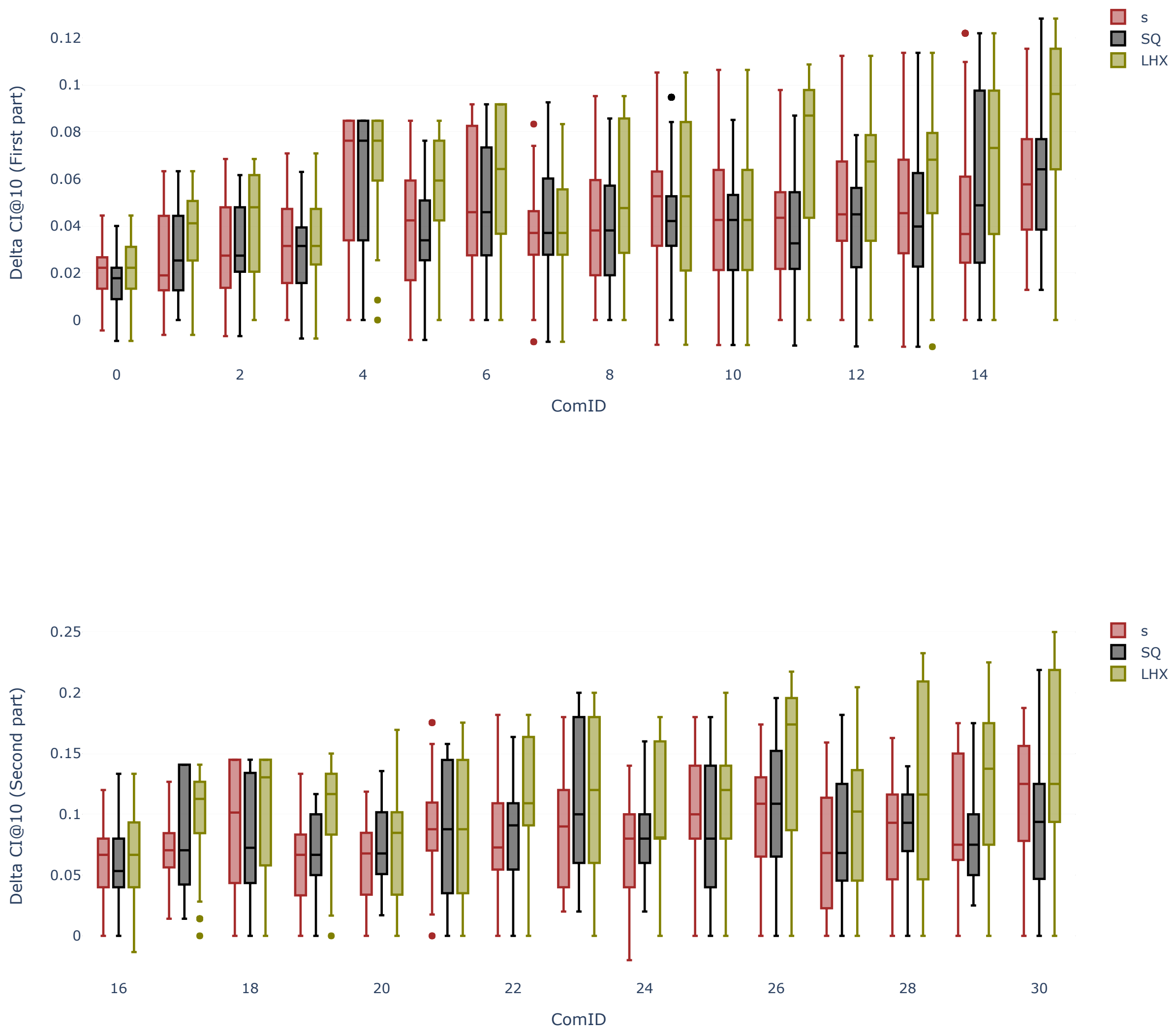


Figure 6.24: Delta CI@10 for Mod-comm

Figures 6.23 and 6.24 illustrate the Delta CI@10 of the methods that consider social structure (namely S, SQ and LHX) by comparing them to the MF baseline method without social structure. Figures 6.25 and 6.26 display the same comparison with the difference in the parameter setting: Delta CI is set at 20. The Delta CI@50 is depicted in the Figures 6.26 and 6.28. An increase is observed in community members after integrating Social regularization methods. There is a significant difference when

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

comparing recommenders with the social structure against the MF baseline method. Especially the LHX method has a significant higher Delta CI in comparison to the Delta CI of the S and SQ methods. This evidence is found for both of the community methods.

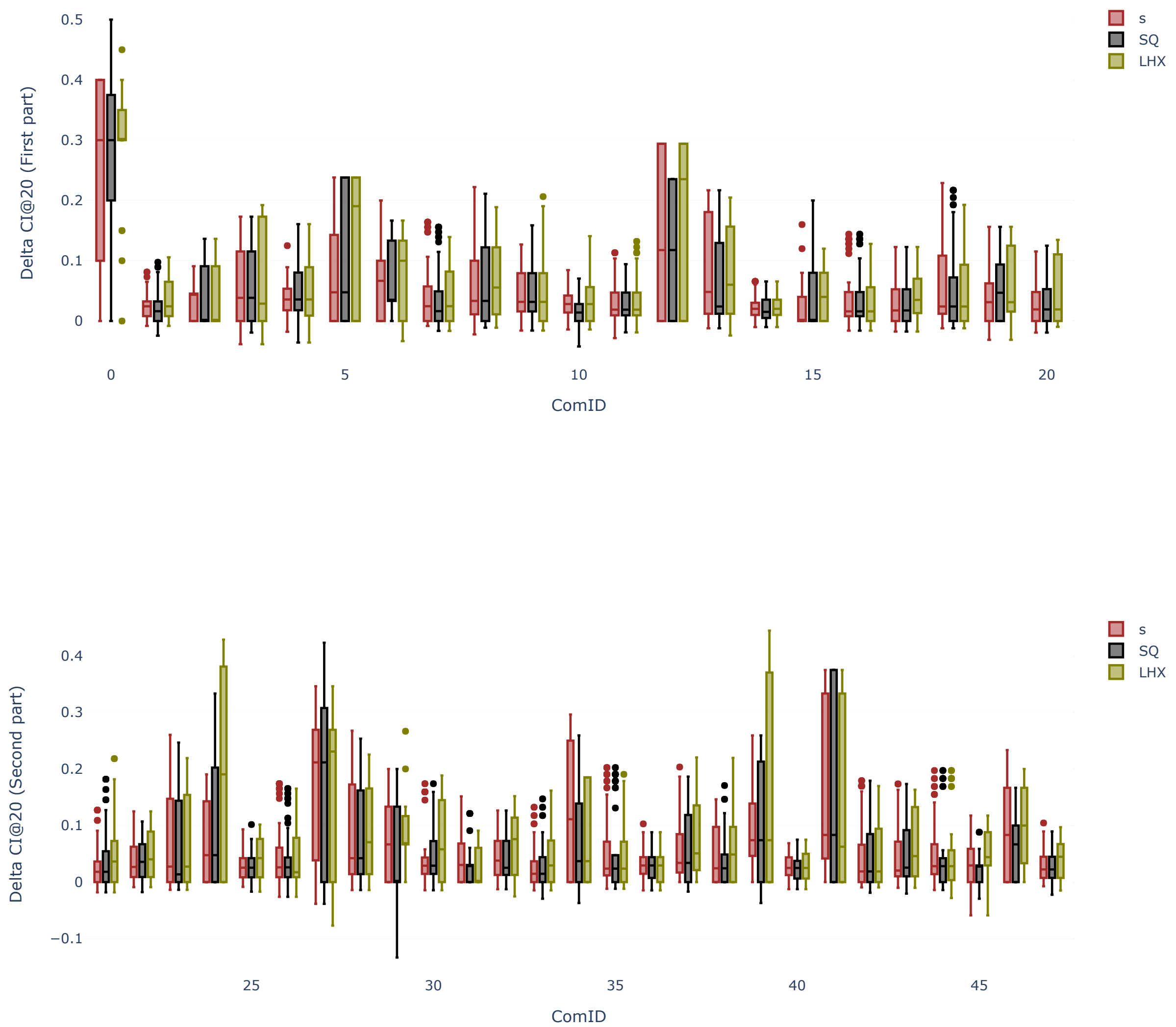


Figure 6.25: Delta CI@20 for Inf-comm

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

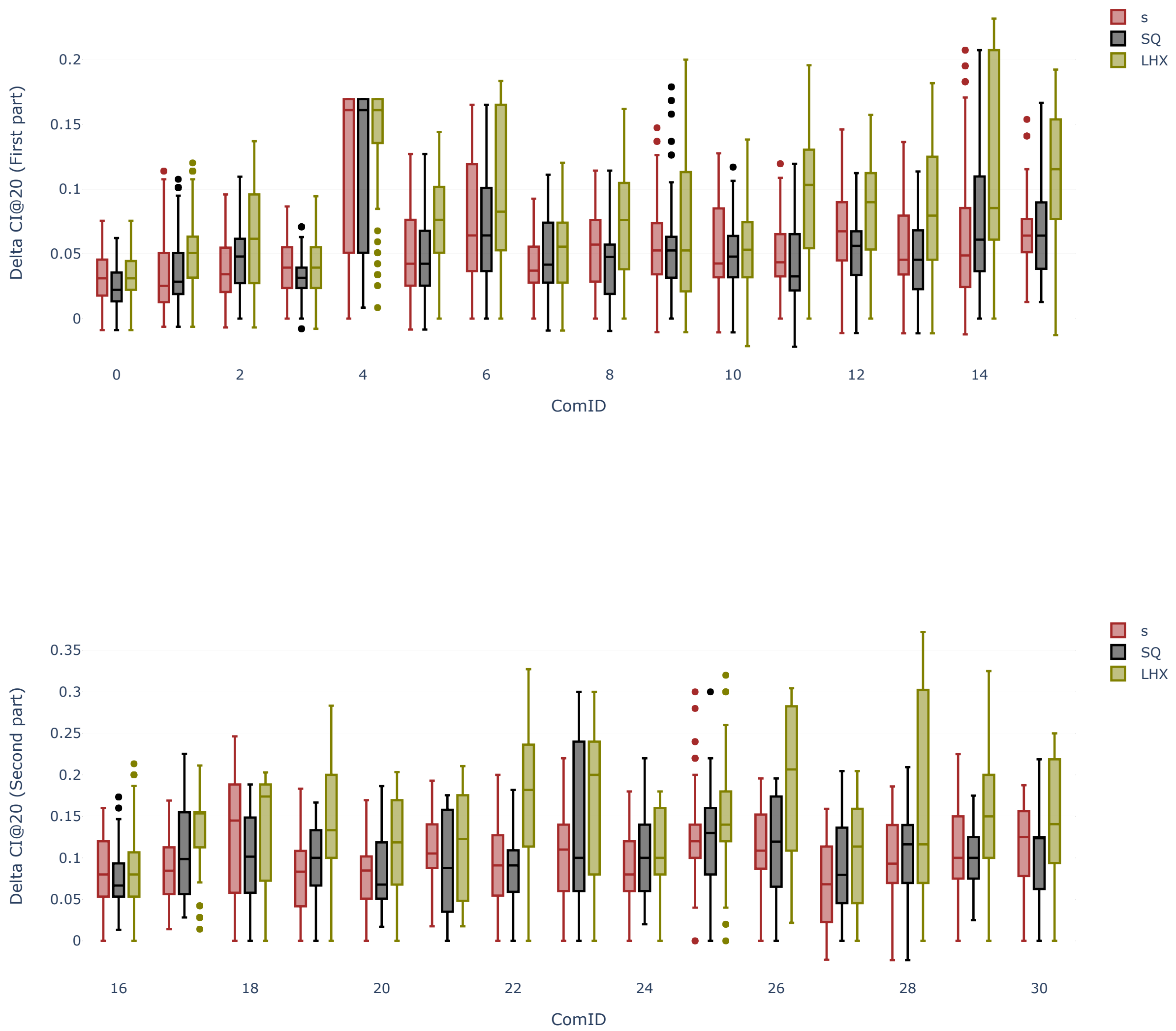


Figure 6.26: Delta CI@20 for Mod-comm

Apart from being suitable to capture the degree of change, the Delta CI is a good measure of the CI impact on users. In real life, some users can be extremely influenced either positively or negatively.

- *Positive* influence is found when it is easy for a user to cope or adapt and enjoy the new environment or community. To detect these users we check where the Delta

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

CI of community members increases extremely. For example, see Figure 6.25 and Figure 6.26. The Delta CI@20 of the S, SQ and LHX recommender methods has many positive outliers, mainly for the S and SQ methods.

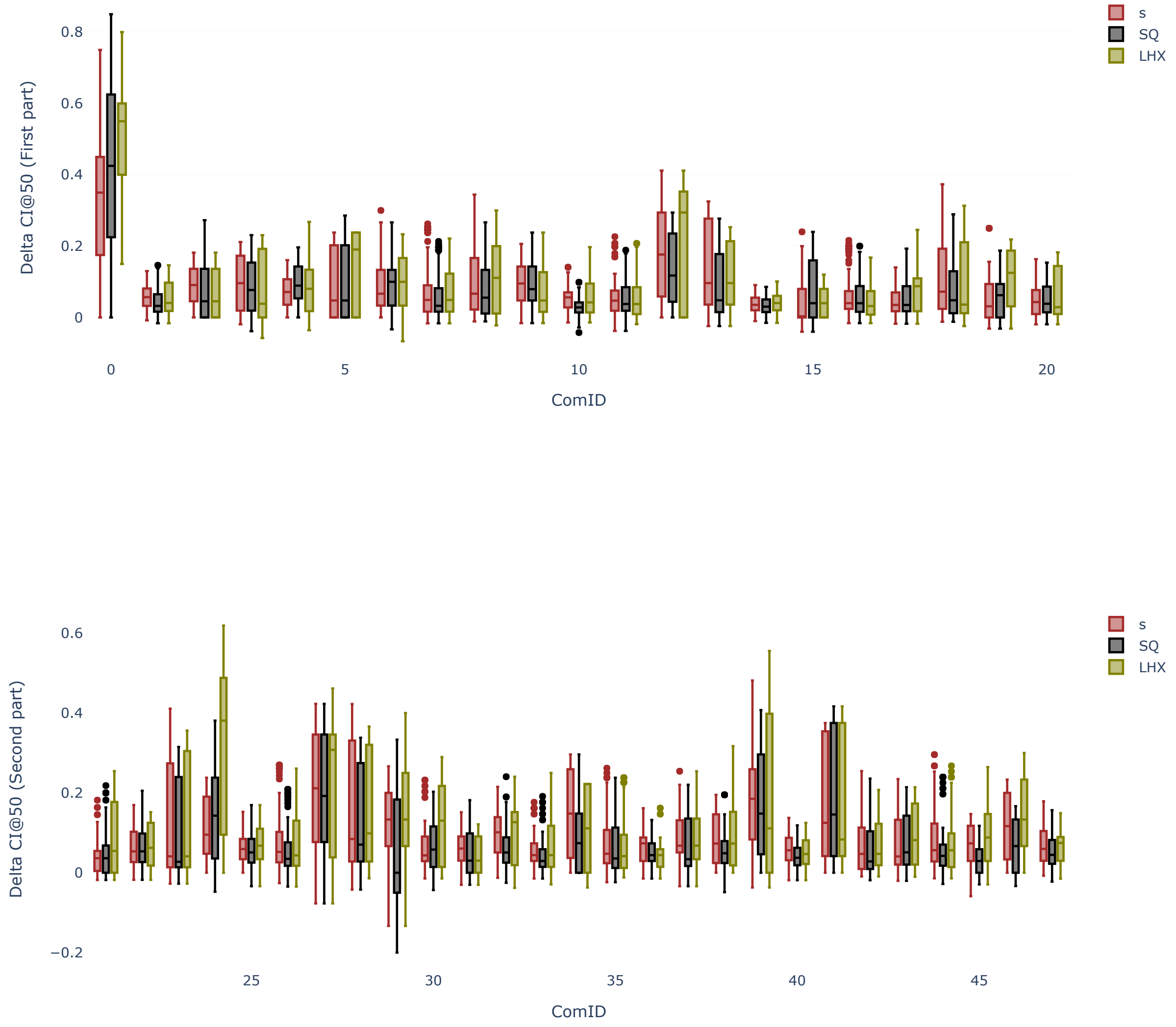
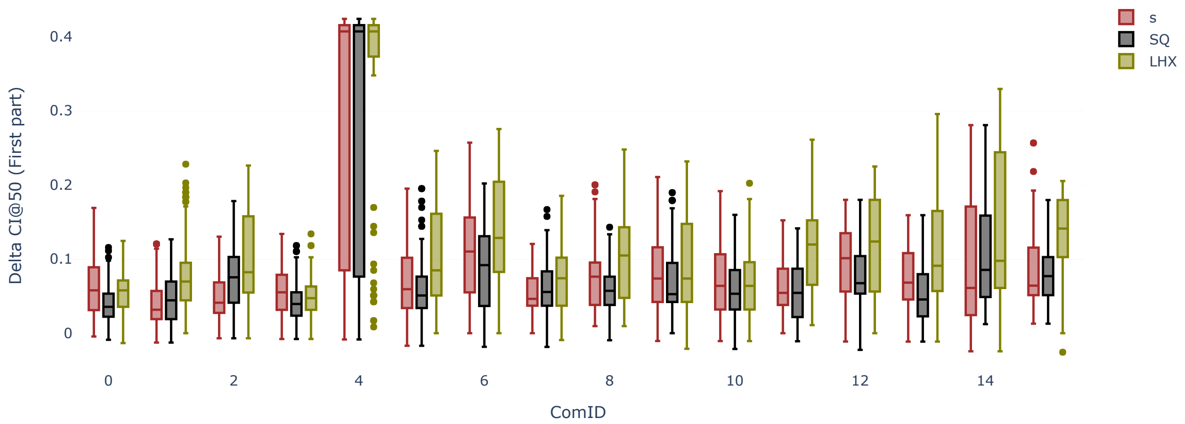


Figure 6.27: Delta CI@50 for Inf-comm

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



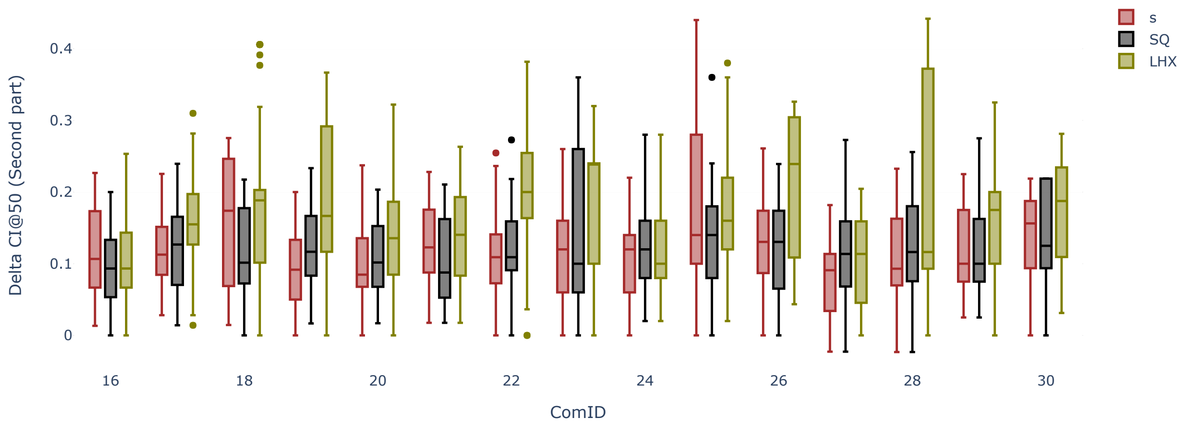


Figure 6.28: Delta CI@50 for Mod-comm

- *Negative* influence is found when it is hard for a user to cope or adapt fast to the new environment or community. To discover these users we examine where the Delta CI of community members decreases extremely or tends to have negative values. For example, see Figure 6.23. As for the Delta CI@10 for the Influencer based communities, we notice two negative outliers of the S method for the 4th community (in brown) and of the SQ method for the 10th and 29th communities

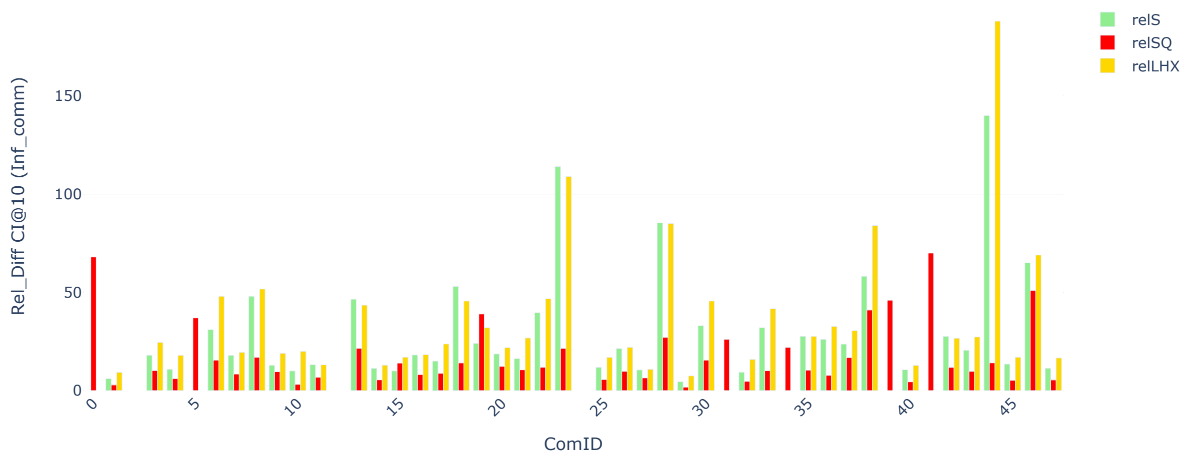
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

(in black). Another example is the Figure 6.24. Delta CI@10 for the Modularity based communities shows two negative outliers of the LHX method for the 4th community (in olive). The main observation is that the number of the negative outliers is relatively small compared to the number of the positive outliers for both community methods.

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

6.4.3 Community Influence for Community View

Relative Difference or Relative Delta (Δ)



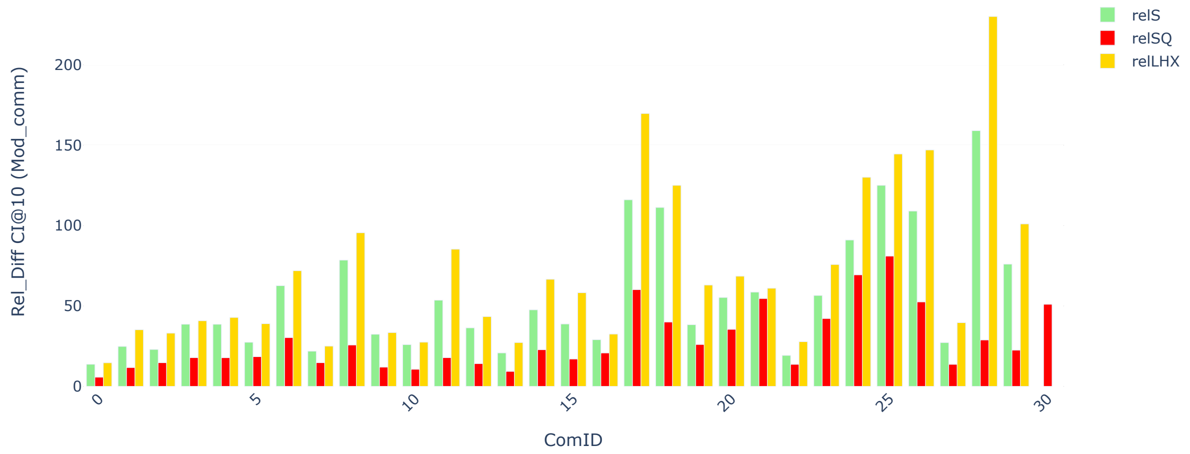


Figure 6.29: Rel-Diff CI@10: Inf-comm and Mod-comm

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

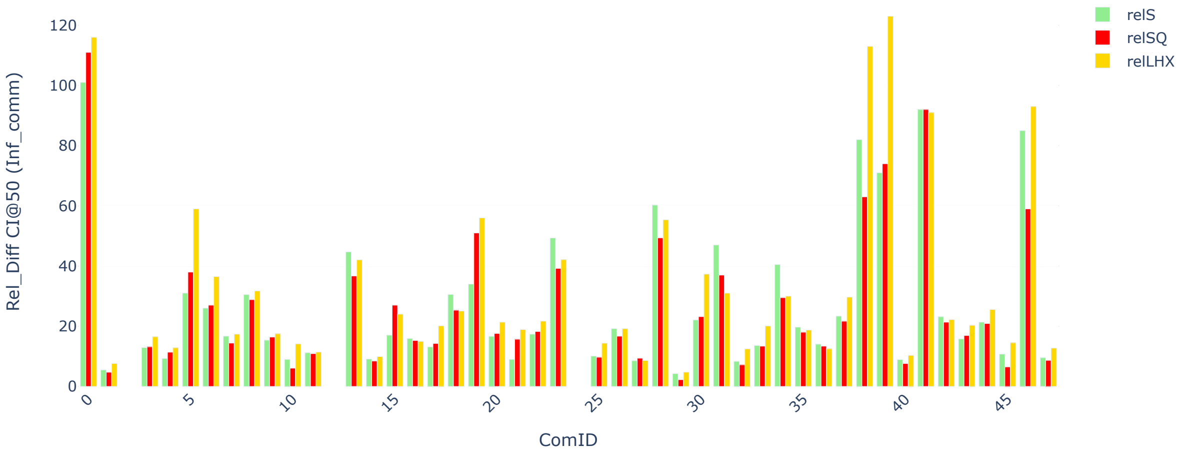
Here our aim is to get a direct insight into the true scale of difference between our recommender methods and the MF baseline when a community influence view is considered.

Figure 6.30: Rel-Diff CI@20 for Inf-comm and Mod-comm

Furthermore, we want to understand at what rate the community influence behavior of

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

the S, SQ and LHX methods changes with respect to the MF baseline. For these purposes, the relative delta of the community influence of a given method X ($r\Delta CI_X@k(c)$) is computed.



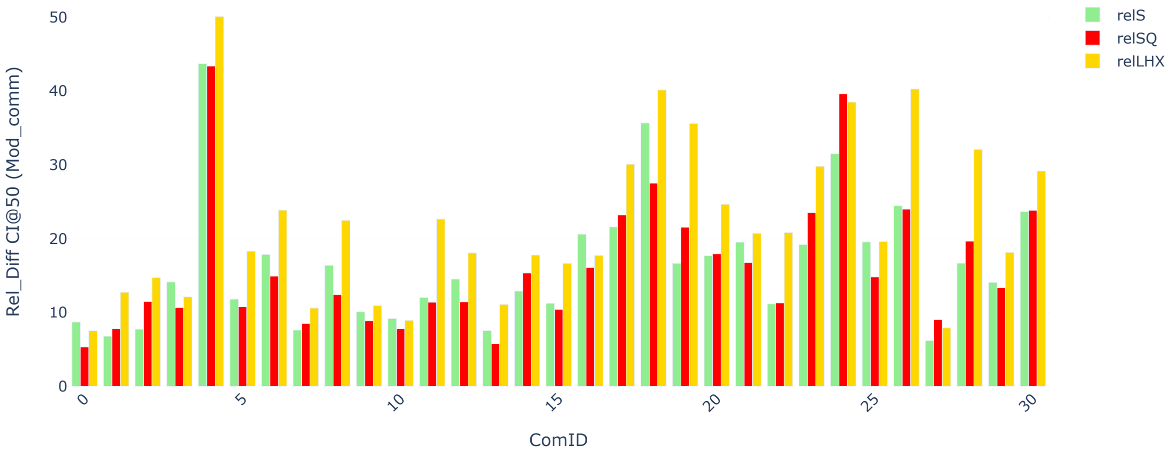


Figure 6.31: Rel-Diff CI@50 for Inf-comm and Mod-comm

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

The relative delta is determined for both, Influencer based community and Modularity based community at the k value equals to 10, 20 and 50. Figure 6.29 displays the difference of each of the $r\Delta CI_S@k(c)$, $r\Delta CI_{SQ}@k(c)$ and $r\Delta CI_{LHX}@k(c)$ methods for CI@10 compared to the MF baseline method. The relative difference varies a lot from one method to another. The important point is that significant difference is observed for each method in comparison to the MF baseline. The maximum relative delta value lies above 200% and the minimum value is around 10%. Since it would be unreasonable to draw firm conclusions from the results of a single value of k, we consider other values of k as well. Figure 6.30 shows the relative differences of the $r\Delta CI_S@k(c)$, $r\Delta CI_{SQ}@k(c)$ and $r\Delta CI_{LHX}@k(c)$ methods for CI@20 compared to the baseline. The barplots for CI@20 display approximately the same trend as the results for CI@10. The last Figure 6.31 illustrates the comparison of the $r\Delta CI_S@k(c)$, $r\Delta CI_{SQ}@k(c)$ and $r\Delta CI_{LHX}@k(c)$ methods for CI@50 compared to the baseline. The barplots again display a significant difference of the methods in consideration in comparison to the MF method, especially for LHX method. That brings us to the following conclusion: the community view based CI for recommenders with social structure is distinguishable from the CI of the recommenders without social structure, such as our baseline MF method.

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

Average and Variance

Statistically, the average value sums up all individual values divided by the number of individuals. To seize that within communities, we compute the average CI@k for both, Influencer based communities and Modularity based communities.

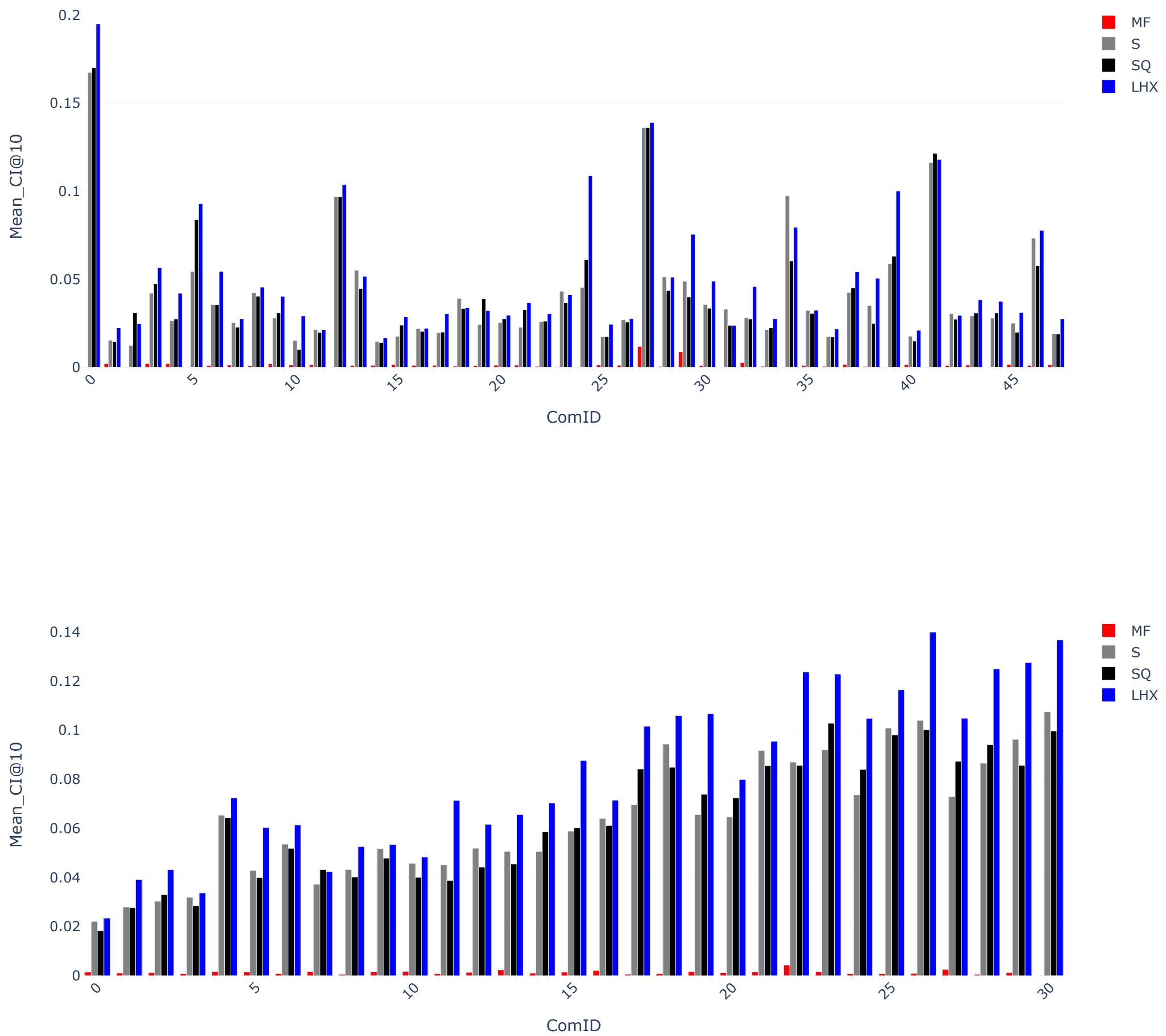


Figure 6.32: Mean CI@10 for Inf-comm and Mod-comm

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

In addition to the average, we calculate the variance of CI@k as it is a very important measure of how far the data points deviate from the mean by looking at the distribution of the data points.

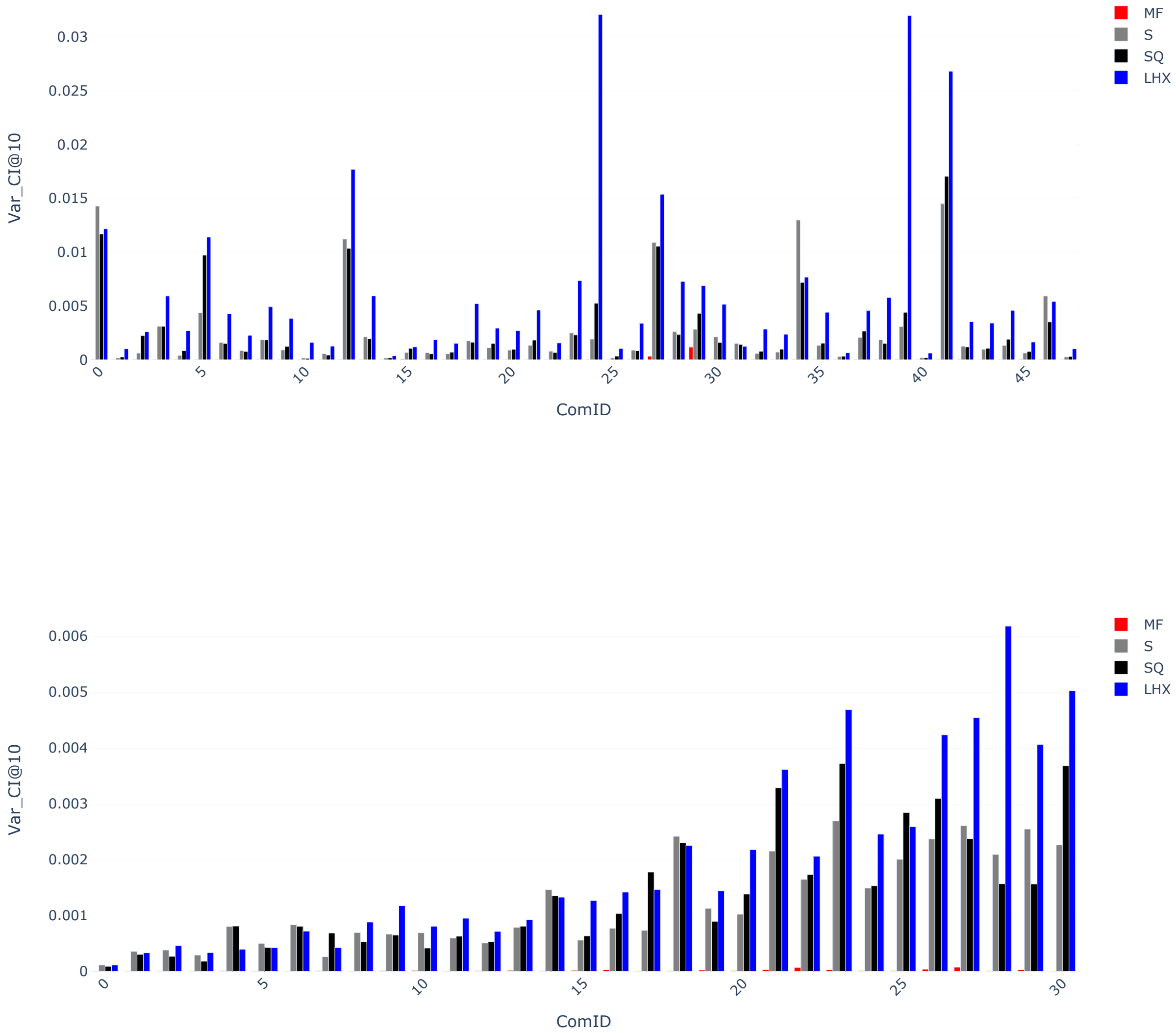


Figure 6.33: Var CI@10 for Inf-comm and Mod-comm

The average and variance are determined for every community and every k value (10, 20 and 50). Figure 6.32 plots the average of CI@10 for all recommender methods (S,

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

SQ, LHX and MF) while Figure 6.33 plots the variance of CI@10. Regarding the MF baseline, the average and the variance CI@k are nearly zero.

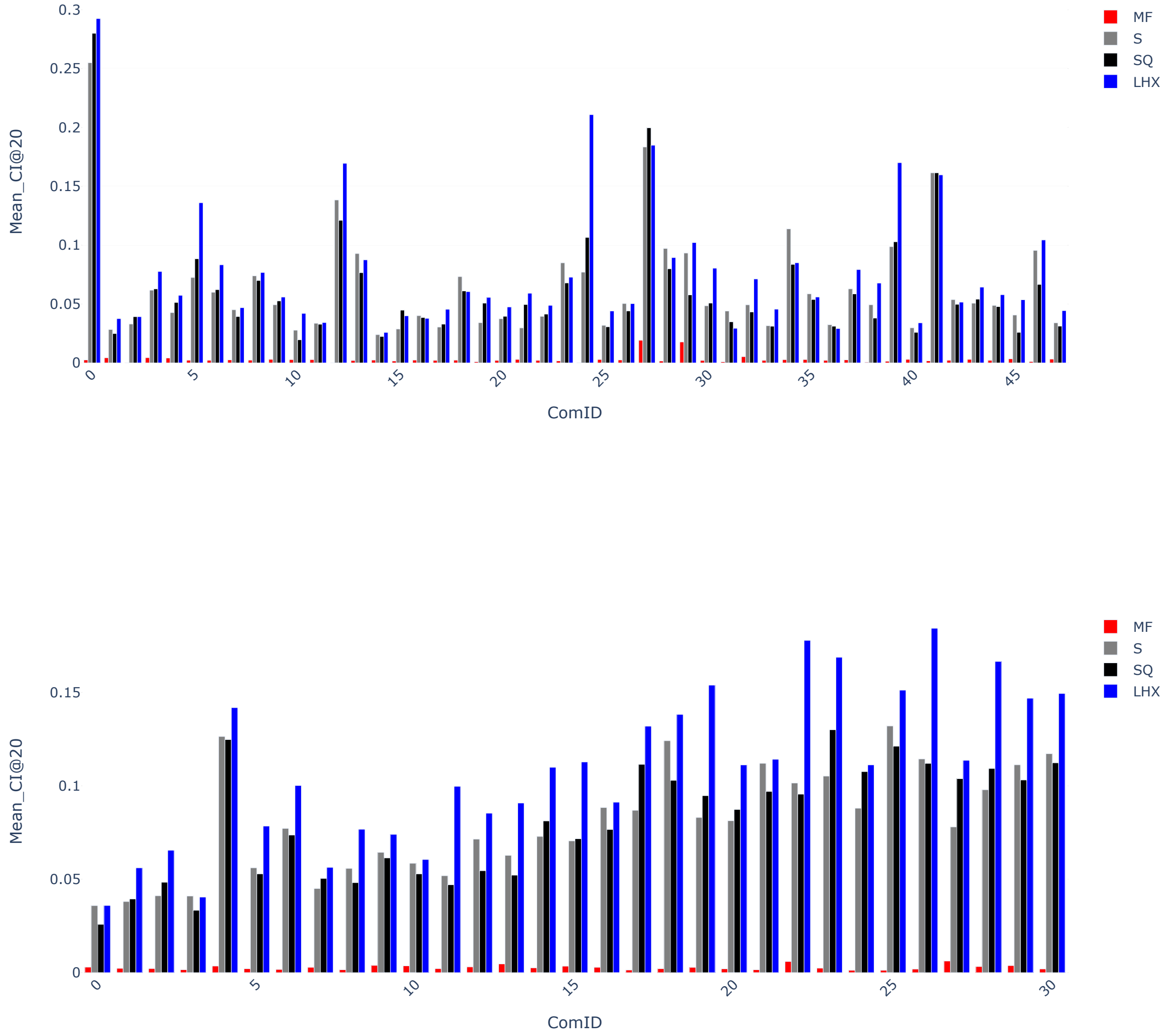
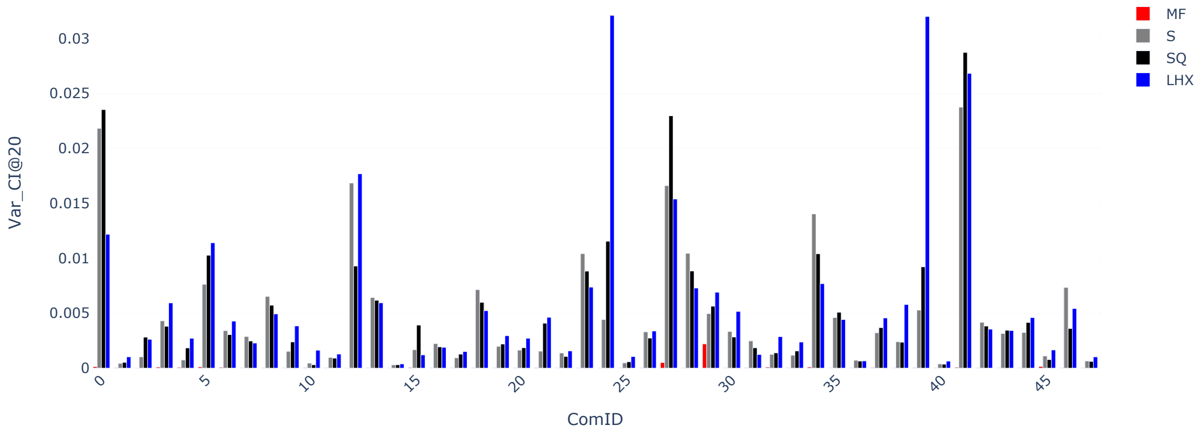


Figure 6.34: Mean CI@20 for Inf-comm and Mod-comm

Barplots for both the average and variance of communities show that the mean and the variance values for the Influencer based communities vary a lot within communities. For example, Figure 6.33 illustrates this for the communities 24 and 39. Unlike the Influencer

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

based communities, the increase is moderate for the Modularity based communities.



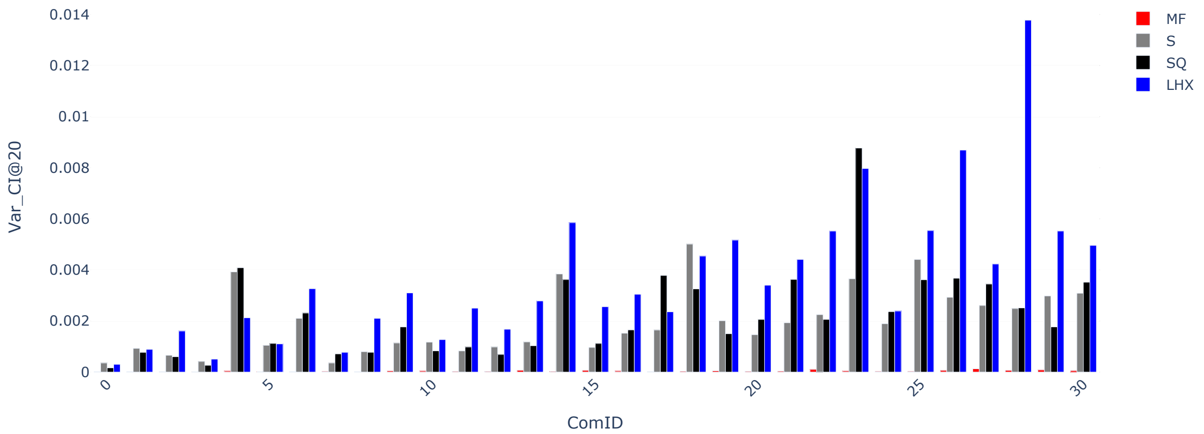


Figure 6.35: Var CI@20 for Inf-comm and Mod-comm

Figure 6.34 plots the average of CI@20 for every community and for all recommender methods (S, SQ, LHX and MF), while Figure 6.35 plots the variance of CI@20 for every community and for all recommender methods (S, SQ, LHX and MF). Comparing the average and the variance of CI@20 to the average and variance of CI@10, we can ascertain

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

the fact that with the increase of k both the mean and variance increase as well.

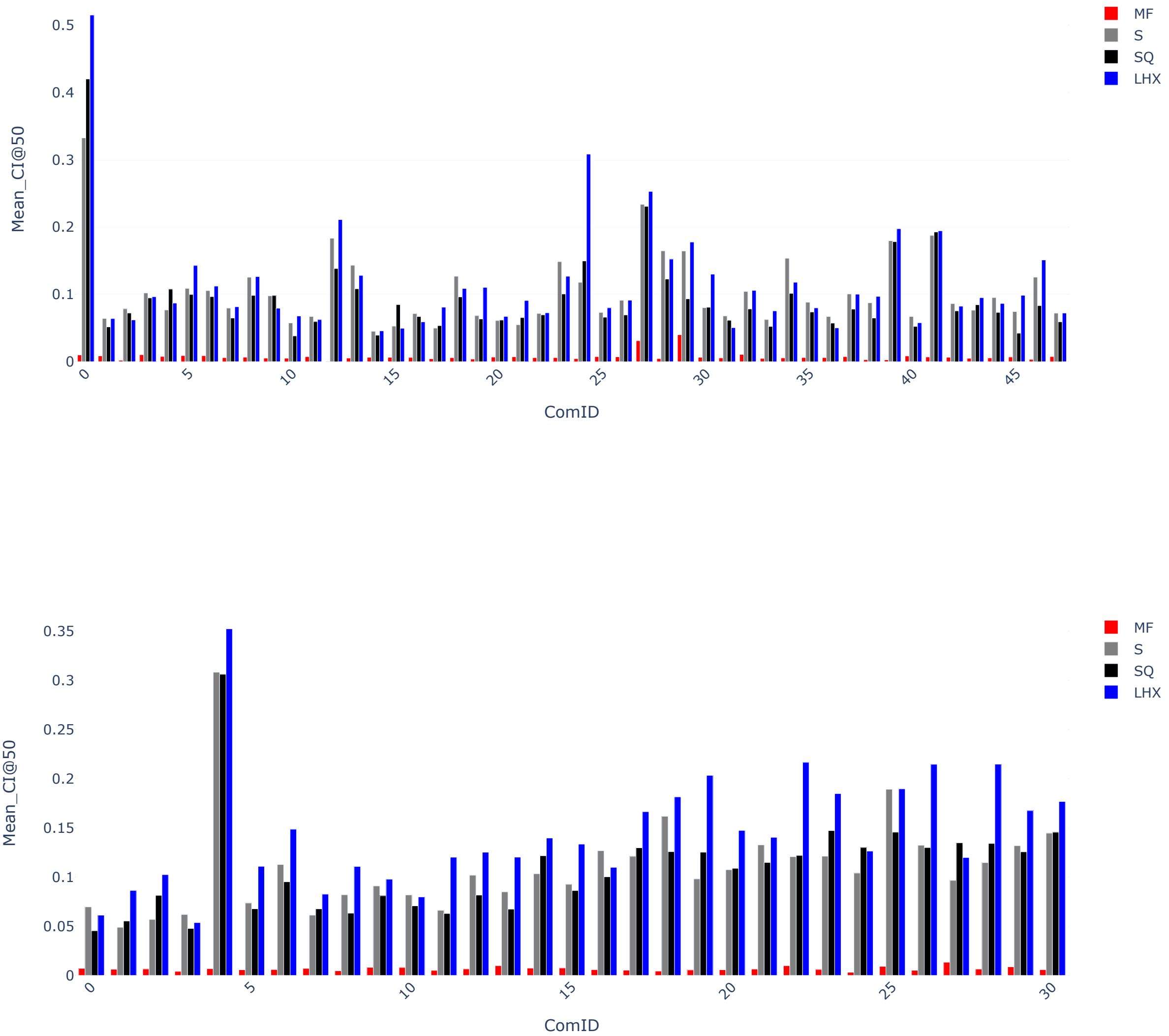
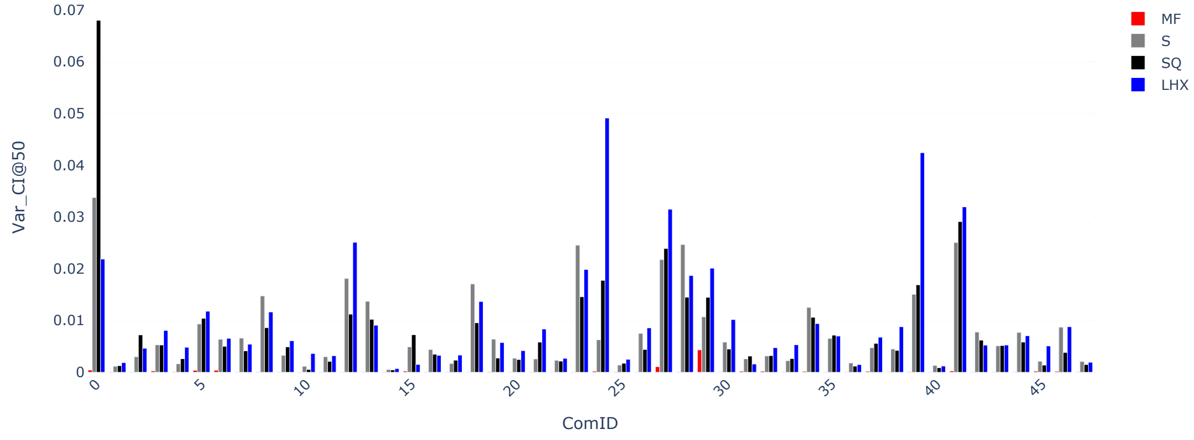


Figure 6.36: Mean CI@50 for Inf-comm and Mod-comm

Figure 6.36 presents the average, while Figure 6.37 presents the variance of the CI@50. We notice that compared to the average and variance of the CI@10 and CI@20, both the mean and variance of the CI@50 increase their values.

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



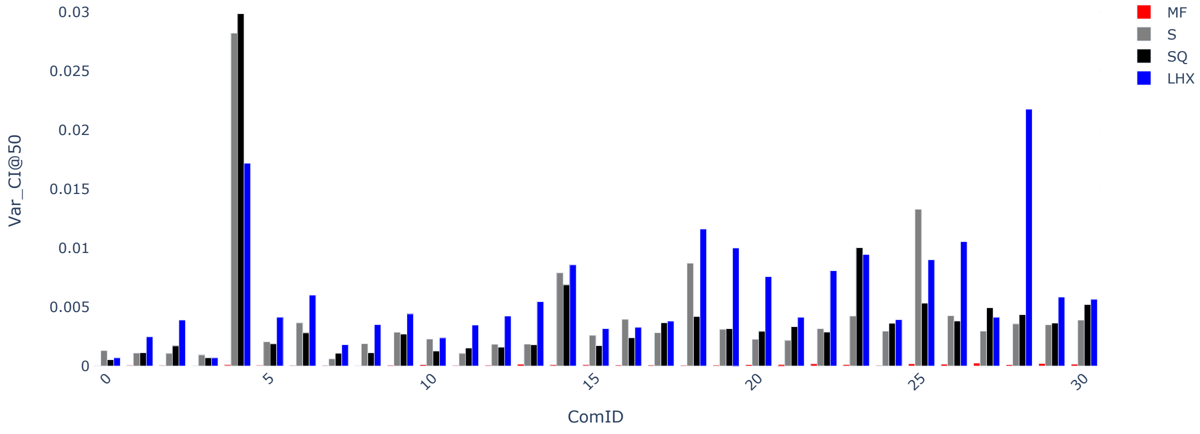


Figure 6.37: Var CI@50 for Inf-comm and Mod-comm

From the obtained results we can clearly see that the variance is particularly small across all CI@k. To sum up, users within the communities strongly influenced by the community behavior, in various degrees.

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

Outliers of Delta (Δ)

For each method, outliers show how many users are extremely positively or negatively influenced by the social community compared to the MF baseline.

Figure 6.38: Outliers CI@10 for Inf-comm and Mod-comm

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

In our case there is an insignificant number of the negative outliers at the low end of the box. Recall that negative outliers can be interpreted as users who find it hard to cope or adapt fast to the new environment or community. We detect these users by looking at the extreme low values of the Delta CI of community members.

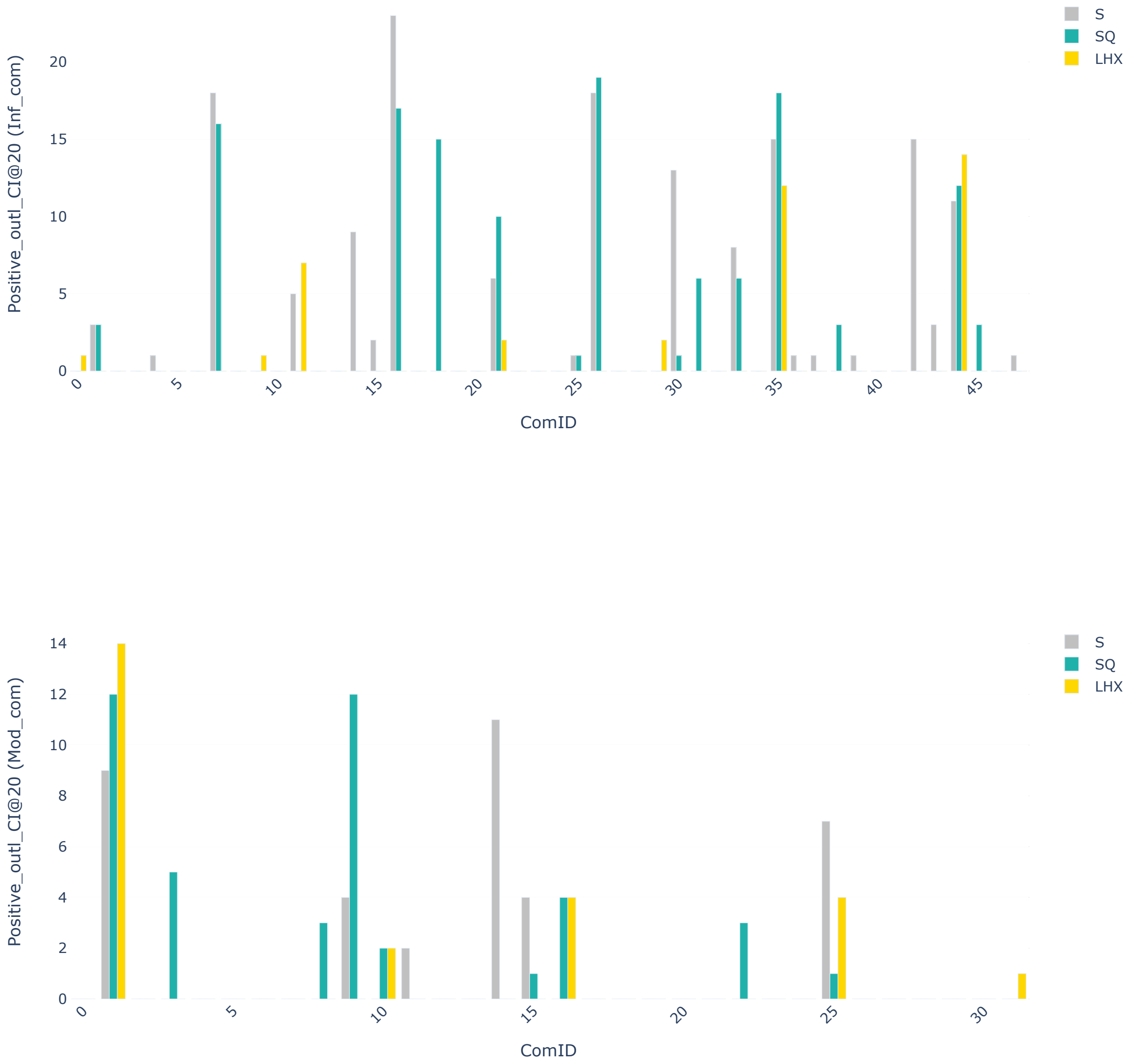
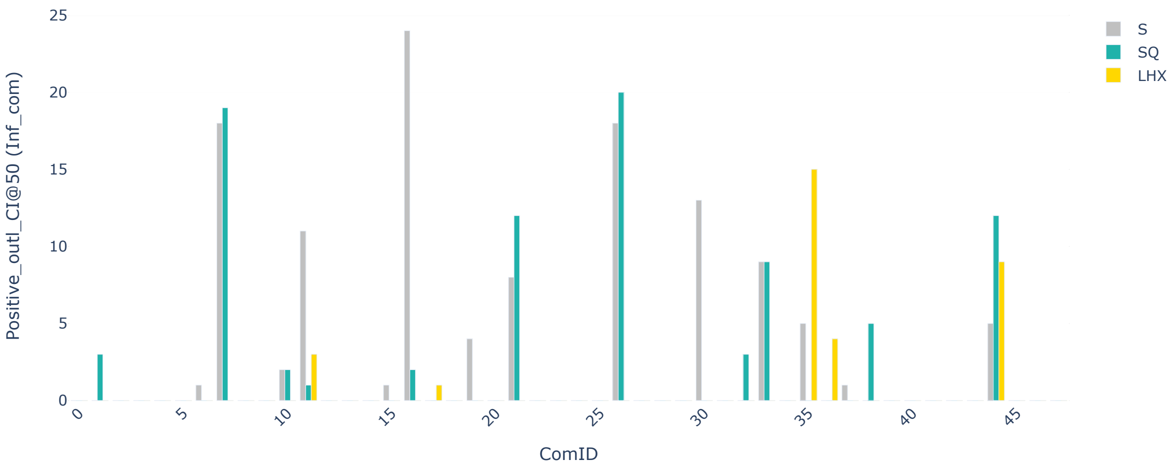


Figure 6.39: Outliers CI@20 for Inf-comm and Mod-comm

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

Figure 6.23 presents results of the Delta CI@10 for the Influencer based communities that show three negative outliers. One negative outlier of the S method for the 4th community (in brown) and two negative outliers of the SQ method for the 10th and 29th communities (in black). These findings correspond to the ones mentioned before.



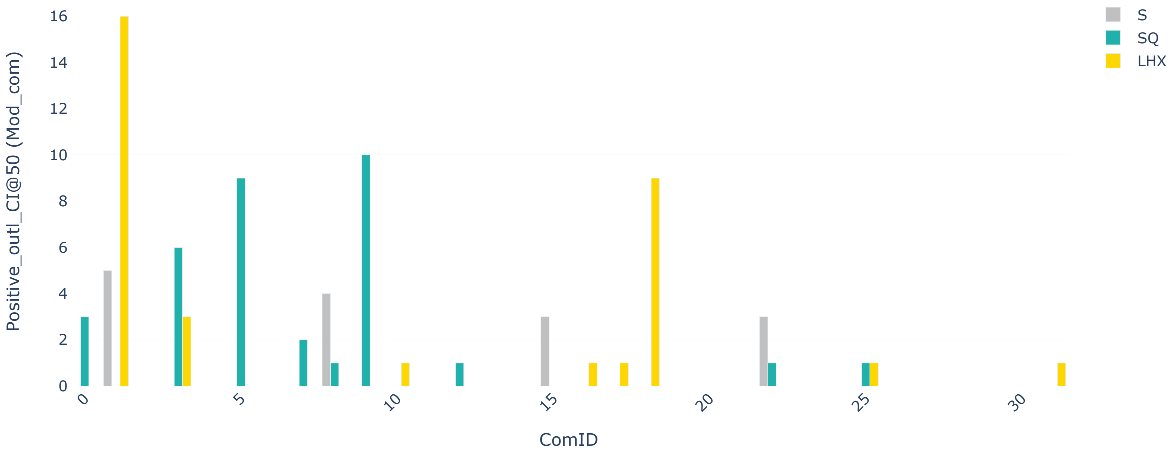


Figure 6.40: Outliers CI@50 for Inf-comm and Mod-comm

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

Another example is Figure 6.24 of the Delta CI@10 for Modularity based communities which shows two negative outliers of the LHX method for the 4th community (in olive).

Users representing positive outliers find it quite easy to cope or adapt and enjoy the new environment or community. There is a noticeable number of the positive outliers at the upper end of the box. For example, examining Figure 6.25 and Figure 6.26 of the Delta CI@20 for the S, SQ and LHX recommender methods, we notice many positive outliers, particularly for the S and SQ methods.

Next, we calculated the number of outliers for every community method (Influencer based community and Modularity based community) and at every k value (10, 20 and 50). Figure 6.38 displays the results for CI@10. There is a remarkable difference between the number of outliers found in Influencer based communities and in Modularity based communities. Influencer based communities have a higher number of outliers. Figure 6.39 shows the number of outliers for CI@20 and Figure 6.40 indicates the number of outliers for CI@50. The barplots show that the number of outliers increases with the increase of the k value for both community methods. Furthermore, the results indicate that the S and SQ methods have a much higher number of outliers compared to the LHX method.

In support to the barplots, we create a comparison table for all social structure recommenders and community methods. Table 6.3 and Table 6.4 show how differently community and social structure methods affect users. For example, in both tables, we can see that for both the Influencer and Modularity based communities the S and SQ methods have a remarkable number of the outliers, while the LHX method has a very few outliers. In conclusion, we can say that the LHX method treats members of the community more fairly.

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

Table 6.3: Outliers (Inf-comm)

ComID	nn	SQ	S	LHX
0	20	0	0	1
1	123	14	3	0
2	22	2	5	0
3	52	0	9	0
4	56	4	3	0
5	21	0	0	0
6	30	0	1	0
7	122	35	36	0
8	90	0	0	0
9	63	1	1	1
10	71	3	2	0
11	106	3	24	18
12	17	0	2	0
13	83	0	0	0
14	198	0	9	0
15	25	0	3	0
16	125	36	47	0
17	57	0	0	1
18	83	15	0	0
19	32	0	8	2
20	104	0	16	0
21	55	22	24	2
22	112	6	1	0
23	73	0	0	0
24	21	0	0	0
25	118	3	1	0
26	115	39	36	0
27	26	0	0	0
28	71	0	0	0
29	15	1	0	2
30	69	12	38	0
31	33	10	0	4
32	79	5	0	0
33	68	21	24	0
34	27	5	0	0
35	84	31	20	27
36	68	1	2	4
37	59	0	3	0
38	41	14	0	0
39	27	0	1	0
40	160	0	0	0
41	24	0	0	0
42	106	0	15	0
43	98	0	3	0
44	71	36	27	37
45	34	6	0	0
46	30	0	0	0
47	134	2	1	0
		327	365	99

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

Table 6.4: Outliers (Mod-comm)

ComID	nn	SQ	S	LHX
0	225	3	0	0
1	158	12	14	30
2	146	0	0	0
3	127	11	0	3
4	118	0	0	0
5	118	9	0	0
6	109	0	0	0
7	108	2	1	0
8	105	4	4	0
9	95	34	4	0
10	94	2	0	3
11	92	0	2	0
12	89	1	0	0
13	88	0	0	0
14	82	0	23	0
15	78	2	7	0
16	75	4	0	5
17	71	0	0	1
18	69	0	0	9
19	60	0	0	0
20	59	0	0	0
21	57	0	5	0
22	55	4	3	0
23	50	0	0	0
24	50	0	0	0
25	50	2	7	5
26	46	0	0	0
27	44	0	0	0
28	43	0	0	0
29	40	0	0	0
30	32	0	0	0
31	11	0	0	2
		90	70	58

6.5 Conclusions

Community members have a high MF similarity, particularly, when these members are friends or have strong social connections. Different social structure recommenders increase the similarity of users. We observe a high number of positive outliers representing users positively affected by the community behavior. This leads us to the fact that users within a community are strongly affected by its behavior. At the individual view with regard to their social circle (i.e., the communities they belong to), our analysis shows that various social recommenders differ in their impact on individuals' preferences, although they have comparable effectiveness.

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

CHAPTER 7

Conclusions

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

7.1 Summary

The summary of the research is presented on the same levels as the research questions:

At the Individual level, we sought to examine whether there was a relationships between social connections and rating behavior in social recommenders. We analysed the popularity of a user based on the centrality measures and the PageRank algorithm in order to understand the network structure and the level of the rating behavior of a user (activity is based on a number of ratings). We found out that the number of ratings made by a user and her centrality in social network were related, particularly when the latter was measured by the number of social connections.

At the Pairwise level, the analysis was built upon the individual level results. Here we examined pairs of users in a more detailed way. We obtained a rating behavior of a user by computing several similarity measures between pairs of users. We then quantified the relation between the rating behavior and the social structure which was measured by several network similarity metrics. We analysed the correlation between both similarities in different views. As a result, rating behavior and social structure seemed to be related. Specifically, rating behavior tended to determine social structure more precisely than the other way around. In addition, we considered pairs of highly popular and unpopular users in a social network and examined the influence in their rating behavior. We also investigated whether friends with similar ratings have a leader-follower relationship in a social network. Concerning the highly popular and unpopular user analysis, we found out that if a user with low social activity was connected with a user with high social activity, we expected their feed-back similarity of RATE-PCC to be almost two times as high as similarity of other pair of friends. Although we could not be certain about the direction of influence, we conjectured that it flowed from a more socially active user to a less active one.

At the Community level, we adopted the findings from the Pairwise level where we found that rating behavior was mostly defined by the type of connection a user had in social activity. The results showed that a popular user when having a high degree centrality in the network could impact the rating behavior of a user who was not active in the network, and vice versa. We dug deeper by implementing social structure recommender models and taking into account different types of users pairs based on their centrality. We extended our comparison to groups of users extracted by community detection mechanisms and by ratings-based neighborhood approaches (i.e., matrix factorization based similarity). The aim was to figure out the following: Whether community members behaved similarly in terms of rating, whether users enjoyed the community or not and whether there were some users that were affected by community or social structure recommenders in an extreme way.

We found out that community members had a high rating similarity, especially when these members had strong social connection. Regarding recommenders, social structure methods increased the similarity of users, and users within a community were strongly affected by the community behavior. We noticed a remarkable number of positive outliers which represented users who were affected positively by the community behavior.

To sum up, there existed a significant number of connections between the rating behavior and the social structure in social recommenders at all levels (among individuals, among pairs of friends, and within communities). Although various social recommenders had comparable effectiveness, they differed in their impact on individuals' preferences, e.g. outliers. Based on these results, we proposed a social recommender that was as effective as existing techniques but treated users more fairly.

7.2 Limitations

7.2.1 Research Limitations

The advantage of using historical data is that it provides us with a better representation of preferences and future behaviors of users compared to the information that we obtain from a self-opinionated user survey or interview. Apparently, we as humans are far away from being able to predict our own behavior. Moreover, historical data also measures user activities and preferences quite accurately, but the question is: *Can the analysis results obtain from analysis experiments based on historical data correctly predict real-world field behavior?*. This brings us to an important limitation of the inability to verify our results in a real experiment scenario.

7.2.2 Open Data Limitations

One of the most important findings of our work is that pairs of type low-high centrality in terms of social connection exhibit stronger correlations in their rating behavior. This means that there is a stronger force of influence between them. Although the direction of

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

the force cannot be identified using the data available, we conjecture that popular users are the ones who exert influence on unpopular ones.

Note that the use of publicly available datasets entails certain limitations. For example, In particular, datasets may have issues with privacy policy, such as trust information. The FilmTrust and CiaoDVD datasets have trust privacy with the weight of trust kept private because of privacy which could lead to quality barriers while doing analysis related to trust in the networks.

7.3 Future work

It could be interesting to consider the use of our findings in different domains and to test them with an updated dataset. For example, health recommender system (HRS) has become a very important platform for healthcare services in today's world when it comes to establishing diagnosis or predicting future behavior of a certain disease. Normally, a large amount of patient data needs to be thoroughly analysed in order to capture patients lifestyle (behavior), social activities, etc. and then promptly adopt the necessary measures.

One possible direction to enhance our approach is to experiment recommender models. For example, regularizers based on hybrid feedback can be used from ratings (explicit feedback) and browsing history, purchase logs (implicit feedback).

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

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

List of Figures

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

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

List of Tables

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

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

Bibliography

[ABS+12] Luca Maria Aiello, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Friendship prediction and homophily in social media. *TWEB*, 6(2):9:1–9:33, 2012.

[AGMcC07] Ioannis Antonellis, Hector Garcia-Molina, and Chi chao Chang. Simrank++: Query rewriting through link analysis of the click graph. Technical Report 2007-32, Stanford InfoLab, 2007.

[AT05] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.*, 17(6):734–749, 2005.

[AT15] Gediminas Adomavicius and Alexander Tuzhilin. Context-aware recommender systems. In Ricci et al. [RRS15], pages 191–226.

[AV16] Shaikhah Alotaibi and Julita Vassileva. Personalized recommendation of research papers by fusing recommendations from explicit and implicit social network. In *Proceedings of the 24th ACM Conference on User Modeling, Adaptation and Personalisation (UMAP 2016)*, 2016.

[BAX10] Halil Bisgin, Nitin Agarwal, and Xiaowei Xu. Investigating homophily in online social networks. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010, Toronto, Canada, August 31 - September 3, 2010, Main Conference Proceedings*, pages 533–536, 2010.

[BOH15] Robin Burke, Michael P. O'Mahony, and Neil J. Hurley. Robust collaborative recommendation. In Ricci et al. [RRS15], pages 961–995.

[Bur02] Robin D. Burke. Hybrid recommender systems: Survey and experiments. *User Model. User-Adapt. Interact.*, 12(4):331–370, 2002.

[Bur12] Robin Burke. *Recommender Systems: An Introduction*, by dietmar jannach, markus zanker, alexander felfernig, and gerhard friedrichcambridge university press, 2011, 336 pages. ISBN: 978-0-521-49336-9. *Int. J. Hum. Comput. Interaction*, 28(1):72–73, 2012.

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

[BV14] Paolo Boldi and Sebastiano Vigna. Axioms for centrality. *Internet Mathematics*, 10(3-4):222–262, 2014.

[CHT15] Shuo Chang, F. Maxwell Harper, and Loren Terveen. Using groups of items for preference elicitation in recommender systems. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1258–1269, New York, NY, USA, 2015. ACM.

[ÇM19] Erion Çano and Maurizio Morisio. Hybrid recommender systems: A systematic literature review. *CoRR*, abs/1901.03888, 2019.

[dABR+14] Guilherme Ferraz de Arruda, André Luiz Barbieri, Pablo Martín Rodriguez, Yamir Moreno, Luciano da Fontoura Costa, and Francisco Aparecido Rodrigues. The role of centrality for the identification of influential spreaders in complex networks. *CoRR*, abs/1404.4528, 2014.

[dMP97] Ramón López de Mántaras and Enric Plaza. Case-based reasoning: An overview. *AI Commun.*, 10(1):21–29, 1997.

[dSPF19] Douglas Veras da Silva, Ricardo B. C. Prudêncio, and Carlos Ferraz. CD-CARS: cross-domain context-aware recommender systems. *Expert Syst. Appl.*, 135:388–409, 2019.

[Dun80] Charles W Dunnett. Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association*, 75(372):796–800, 1980.

[EH05] Allan Ellis and Tatsuya Hagino, editors. *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005*. ACM, 2005.

[EK10] David A. Easley and Jon M. Kleinberg. *Networks, Crowds, and Markets - Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[ELKR11] Michael D. Ekstrand, Michael Ludwig, Joseph A. Konstan, and John Riedl. Rethinking the recommender research ecosystem: reproducibility, openness, and lenskit. In Bamshad Mobasher, Robin D. Burke, Dietmar Jannach, and Gediminas Adomavicius, editors, *Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23-27, 2011*, pages 133–140. ACM, 2011.

[Fre78] Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215 – 239, 1978.

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

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

[GIL+09] Marco De Gemmis, Leo Iaquinta, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, and Giovanni Semeraro. Preference learning in recommender systems. In *In Preference Learning (PL-09) ECML/PKDD-09 Workshop*, 2009.

[Guy15] Ido Guy. Social recommender systems. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 511–543. Springer, 2015.

[GZ16] Wei-Feng Guo and Shao-Wu Zhang. A general method of community detection by identifying community centers with affinity propagation. *Physica A: Statistical Mechanics and its Applications*, 447:508 – 519, 2016.

[GZTY14] Guibing Guo, Jie Zhang, Daniel Thalmann, and Neil Yorke-Smith. ETAF: an extended trust antecedents framework for trust prediction. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2014*, pages 540–547, 2014.

[GZY13] Guibing Guo, Jie Zhang, and Neil Yorke-Smith. A novel bayesian similarity measure for recommender systems. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence IJCAI*, pages 2619–2625, 2013.

[HCZ04] Zan Huang, Hsinchun Chen, and Daniel Dajun Zeng. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Trans. Inf. Syst.*, 22(1):116–142, 2004.

[HKV08] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, pages 263–272, 2008.

[HZ11] Neil Hurley and Mi Zhang. Novelty and diversity in top-n recommendation – analysis and evaluation. *ACM Trans. Internet Technol.*, 10(4):14:1–14:30, March 2011.

[HZC07] Zan Huang, Daniel Dajun Zeng, and Hsinchun Chen. Analyzing consumer-product graphs: Empirical findings and applications in recommender systems. *Management Science*, 53(7):1146–1164, 2007.

[JCL+12] Meng Jiang, Peng Cui, Rui Liu, Qiang Yang, Fei Wang, Wenwu Zhu, and Shiqiang Yang. Social contextual recommendation. In *CIKM*, pages 45–54, 2012.

[JCS04] Rong Jin, Joyce Y. Chai, and Luo Si. An automatic weighting scheme for collaborative filtering. In *SIGIR*, pages 337–344, 2004.

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

[JE10] Mohsen Jamali and Martin Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *RecSys*, pages 135–142, 2010.

[JLZZ09] Ci-Hang Jin, Jian-Guo Liu, Yi-Cheng Zhang, and Tao Zhou. Adaptive information filtering for dynamic recommender systems. *CoRR*, abs/0911.4910, 2009.

[JW02] Glen Jeh and Jennifer Widom. Simrank: A measure of structural-context similarity. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 538–543, New York, NY, USA, 2002. ACM.

[KC08] Won Kim and Hyung-Jin Choi, editors. *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication, ICUIMC 2008, Suwon, Korea, January 31 - February 01, 2008*. ACM, 2008.

[KC18] Yoosun Jamie Kim and Robert A Cribbie. Anova and the variance homogeneity assumption: Exploring a better gatekeeper. *The British journal of mathematical and statistical psychology*, 71 1:1–12, 2018.

[KKLP06] Steffen Kolarczyk, Manuel Koch, Klaus-Peter Löhr, and Karl Pauls. Sectool - supporting requirements engineering for access control. In *Emerging Trends in Information and Communication Security, International Conference, ETRICS 2006, Freiburg, Germany, June 6-9, 2006, Proceedings*, pages 254–267, 2006.

[KWG+12] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Model. User-Adapt. Interact.*, 22(4-5):441–504, 2012.

[LA13] Xin Liu and Karl Aberer. Soco: a social network aided context-aware recommender system. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 781–802, 2013.

[LH16] Philip Lenhart and Daniel Herzog. Combining content-based and collaborative filtering for personalized sports news recommendations. In *Proceedings of the 3rd Workshop on New Trends in Content-Based Recommender Systems co-located with ACM Conference on Recommender Systems (RecSys 2016), Boston, MA, USA, September 16, 2016.*, pages 3–10, 2016.

[LHN06] E. A. Leicht, Petter Holme, and M. E. J. Newman. Vertex similarity in networks. *Phys. Rev. E*, 73:026120, Feb 2006.

[LLDM08] Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th International Conference on World

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

Wide Web, WWW 2008, Beijing, China, April 21-25, 2008, pages 695–704, 2008.

[LNPC15] András London, Tamás Németh, András Pluhár, and T. Csendes. A local pagerank algorithm for evaluating the importance of scientific articles. *Annales Mathematicae et Informaticae*, 44:131–140, 2015.

[LWTM15] Hui Li, Dingming Wu, Wenbin Tang, and Nikos Mamoulis. Overlapping community regularization for rating prediction in social recommender systems. In *RecSys*, pages 27–34, 2015.

[MA07] Paolo Massa and Paolo Avesani. Trust-aware recommender systems. In *RecSys*, pages 17–24, 2007.

[MB15] Kanish Manuja and Arnab Bhattacharya. Using social connections to improve collaborative filtering. In *Proceedings of the Second ACM IKDD Conference on Data Sciences, Bangalore, CoDS 2015, India, March 18-21, 2015*, pages 140–141, 2015.

[MBBW07] Bamshad Mobasher, Robin D. Burke, Runa Bhaumik, and Chad Williams. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Trans. Internet Techn.*, 7(4):23, 2007.

[MCLZ18] Yitong Meng, Guangyong Chen, Jiajin Li, and Shengyu Zhang. Psrec: Social recommendation with pseudo ratings. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, pages 397–401, New York, NY, USA, 2018. ACM.

[MH05] Sung-Hwan Min and Ingoo Han. Detection of the customer time-variant pattern for improving recommender systems. *Expert Syst. Appl.*, 28(2):189–199, 2005.

[MKL09] Hao Ma, Irwin King, and Michael R. Lyu. Learning to recommend with social trust ensemble. In *SIGIR*, pages 203–210, 2009.

[MMC17] Nitin Mishra, Vimal Mishra, and Saumya Chaturvedi. Tools and techniques for solving cold start recommendation. In *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*, IML '17, pages 11:1–11:6, New York, NY, USA, 2017. ACM.

[MSW18] Carine Pierrette Mukamakuza, Dimitris Sacharidis, and Hannes Werthner. Mining user behavior in social recommender systems. In *WIMS*, pages 37:1–37:6. ACM, 2018.

[Muk17] Carine Pierrette Mukamakuza. Analyzing the impact of social connections on rating behavior in social recommender systems. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, UMAP '17, pages 322–326, New York, NY, USA, 2017. ACM.

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

[MYLK08] Hao Ma, Haixuan Yang, Michael R. Lyu, and Irwin King. Sorec: social recommendation using probabilistic matrix factorization. In *CIKM*, pages 931–940, 2008.

[MZL+11] Hao Ma, Dengyong Zhou, Chao Liu, Michael R. Lyu, and Irwin King. Recommender systems with social regularization. In *WSDM*, pages 287–296, 2011.

[New05] MEJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):323–351, 2005.

[New06] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[New16] M. E. J. Newman. Community detection in networks: Modularity optimization and maximum likelihood are equivalent. *CoRR*, abs/1606.02319, 2016.

[NGT12] A. Nazemian, H. Gholami, and F. Taghiyareh. An improved model of trust-aware recommender systems using distrust metric. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1079–1084, Aug 2012.

[OJ06] Gary M. Olson and Robin Jeffries, editors. *Extended Abstracts Proceedings of the 2006 Conference on Human Factors in Computing Systems, CHI 2006, Montréal, Québec, Canada, April 22-27, 2006*. ACM, 2006.

[PBMW99] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web, 1999.

[PM13] Anu Pradhan and G. (Kumar) Mahinthakumar. Finding all-pairs shortest path for a large-scale transportation network using parallel floyd-warshall and parallel dijkstra algorithms. *Journal of Computing in Civil Engineering*, 27(3):263–273, 2013.

[PTG08] Cosimo Palmisano, Alexander Tuzhilin, and Michele Gorgoglione. Using context to improve predictive modeling of customers in personalization applications. *IEEE Trans. Knowl. Data Eng.*, 20(11):1535–1549, 2008.

[QSZL19] Tong Qiao, Wei Shan, Mingli Zhang, and Chen Liu. How to facilitate knowledge diffusion in complex networks: The roles of network structure, knowledge role distribution and selection rule. *International Journal of Information Management*, 47:152 – 167, 2019.

[RIS+94] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *CSCW '94, Proceedings of the Conference on Computer*

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

Supported Cooperative Work, Chapel Hill, NC, USA, October 22-26, 1994, pages 175–186, 1994.

[RKTY10] Bharat Rao, Balaji Krishnapuram, Andrew Tomkins, and Qiang Yang, editors. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*. ACM, 2010.

[RRS15] Francesco Ricci, Lior Rokach, and Bracha Shapira, editors. *Recommender Systems Handbook*. Springer, 2015.

[RS05] Jason D. M. Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *ICML*, pages 713–719, 2005.

[RSK16] Waleed Reafee, Naomie Salim, and Atif Khan. The power of implicit social relation in rating prediction of social recommender systems. In *PloS one*, 2016.

[Saw09] Steven F. Sawyer. Analysis of variance: The fundamental concepts. *Journal of Manual & Manipulative Therapy*, 17(2):27E–38E, 2009.

[Sch15] Daniel Schall. Social network-based recommender systems. In *Springer International Publishing*, 2015.

[Sco00] J. Scott. *Social Network Analysis: A Handbook*. SAGE Publications, 2000.

[SKKR01] Badrul M. Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW*, pages 285–295, 2001.

[SKKR02] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Incremental singular value decomposition algorithms for highly scalable recommender systems. In *Fifth International Conference on Computer and Information Science*, pages 27–28, 2002.

[SLA10] Alan Said, Ernesto W. De Luca, and Sahin Albayrak. How social relationships affect user similarities. In *SRS Workshop*, 2010.

[SM07] Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In *NIPS*, pages 1257–1264, 2007.

[SPUP02] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 253–260, New York, NY, USA, 2002. ACM.

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

[SQBB10] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[TKBK17] Amit Tiroshi, Tsvi Kuflik, Shlomo Berkovsky, and Mohamed Ali Kâafar. Graph based recommendations: From data representation to feature extraction and application. *CoRR*, abs/1707.01250, 2017.

[Tuk49] John W Tukey. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114, 1949.

[TW02] Loren G. Terveen and Dennis R. Wixon, editors. *Extended abstracts of the 2002 Conference on Human Factors in Computing Systems, CHI 2002, Minneapolis, Minnesota, USA, April 20-25, 2002*. ACM, 2002.

[ZKL+10] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matús Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences of the United States of America*, 107(10):4511—4515, March 2010.

[ZXNL15] Shaofeng Zhang, Wei Xiong, Wancheng Ni, and Xin Li. Value of big data to finance: observations on an internet credit service company in china. *Financial Innovation*, 1(1):17, Dec 2015.

[ZYKL17] Huan Zhao, Quanming Yao, James T. Kwok, and Dik Lun Lee. Collaborative filtering with social local models. In Vijay Raghavan, Srinivas Aluru, George Karypis, Lucio Miele, and Xindong Wu, editors, *ICDM*, pages 645–654. IEEE Computer Society, 2017.

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