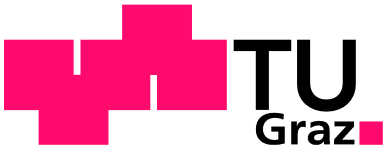
Graz University of Technology
Institute of Information Systems and Computer Media (IICM)
8010 Graz, Inffeldgasse 16c



Predicting Trading Interactions in Trading, Online and Location-Based Social Networks

MASTER THESIS

Lukas Eberhard, BSc

Advisor:

Univ.-Prof. Dipl.-Ing. Dr.techn. Frank Kappe

Head of the Institute of Information Systems and Computer Media

Co-advisor:

Dipl.-Ing. Dr.techn. Christoph Trattner, BSc

Head of the Social Computing Research Group and
Deputy Division Manager of the Knowledge Service Area @ Know-Center

Graz, December 2014

Credits and Copyright

The thesis was written using Arkaitz Zubiaga's
skeleton, which can be downloaded from
`http://www.zubiaga.org/thesis/`.
Many thanks Arkaitz for letting me use this awesome
template for the writing of my master thesis.

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Date | Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe[1].

Datum | Unterschrift

[1] Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Abstract

Link prediction in networks is applied in many places to, for example, compute suggestions for recommender systems or to predict future occurrences based on past events. User attributes or recorded user actions are used to compute a prediction of upcoming interactions. Although the problem of predicting links between users has been extensively studied in the past, research investigating this issue in more than one kind of network is rare. To contribute to this scarce amount of research, the study in this thesis investigated the extend to which trading interactions between sellers and buyers within an online marketplace platform can be predicted based on three diverse network sources – an online social network, a location-based social network and a trading network. To that end two approaches were applied to this problem – supervised learning and unsupervised learning.

The study was conducted in the context of the virtual world Second Life. For that purpose, the data of the online social network of Second Life were crawled, bots automatically monitored user information of the location-based social network of Second Life over a long time period and purchases of the trading network of Second Life were obtained for the experiments. The huge amount of data was analyzed and prepared for further usage. Overall 50 topological and homophilic features were generated which were afterwards used in different constellations to predict trading interactions between user pairs. Supervised and unsupervised learning methods were applied, whereby success rates up to 85.90% could be achieved.

Kurzfassung

Verknüpfungs-Prognosen in Netzwerken werden vielerorts eingesetzt, um beispielsweise Vorschläge für Empfehlungssysteme zu errechnen oder zukünftige Vorgänge anhand vergangener Ereignisse vorherzusagen. Mittels unterschiedlicher Benutzermerkmale oder aufgezeichneten Handlungen wird versucht, zukünftig eintretende Interaktionen vorherzusagen. Obwohl das Thema Link-Prediction-Problem in der Forschung in letzter Zeit intensiv behandelt wurde, ist die Behandlung von nicht nur einem Netzwerk in diesem Zusammenhang noch nahezu unerforscht. Um zu diesem Forschungsgebiet beizutragen, werden die Ergebnisse einer Studie über die Vorhersage von Handelsbeziehungen zwischen Benutzern basierend auf drei unterschiedlichen Datenquellen (soziales Online-Netzwerk, Positions-Netzwerk und Verkaufsplattform) in dieser Arbeit präsentiert. Dazu wurden zwei verschiedene Methoden angewendet – überwachtes und unüberwachtes Lernen.

Die verwendeten Daten in dieser Studie stammen aus der virtuellen Welt Second Life. Zu diesem Zweck wurden die Daten des sozialen Online-Netzwerks von Second Life aufgefasst, Benutzerinformationen dessen Positions-Netzwerks wurden automatisiert durch Bots über längere Zeit gesammelt und Daten der Ein- und Verkäufe der eigenen Second Life-Verkaufsplattform wurden beschafft, um eine Basis für die Experimente in dieser Arbeit zu schaffen. Die enorme Datenmenge wurde analysiert und für die weitere Verwendung vorbereitet. Insgesamt 50 topologische und homophile Merkmale wurden generiert und in unterschiedlichen Konstellationen zur Vorhersage von Handelsinteraktionen zwischen Benutzerpaaren verwendet. Überwachtes und unüberwachtes Lernen wurde bei den Experimenten angewendet, wodurch Erfolgsquoten bis zu 85.90% erreicht werden konnten.

Acknowledgments

I would like to express my gratitude to my advisor, Dr. Frank Kappe, and my mentor, Dr. Christoph Trattner, for their dedicated advice and support. Furthermore, I want to thank Dr. Michael Steurer for introducing me to this topic as well as for his encouraging engagement.

Also, I would like to thank all my colleagues at the Institute of Information Systems and Computer Media at Graz University of Technology for spending their precious time with giving me advice and for the indispensable daily creative and inspiring coffee breaks.

Last, but not least, I would like to thank my family, my girlfriend Melitta and my friends for their unconditional love, steady encouragement and great support during my studies. Without them, I would have never made it this far in life. Thank you!

Contents

List of Figures

List of Tables

1 Introduction

The problem addressed in this master thesis is a particular kind of link prediction problem. The question is to what extent it is possible to predict who will buy from whom and who will sell to whom, or in other words, who will trade with whom in the future. The predictions are based on several sources of data: an online social network, a location-based social network and a trading network.

1.1 Motivation

Social networks offer useful information about the relations between their users and their social characteristics [10]. A huge part of the recent research activity in social networks was the link prediction problem, which was concerned with the forecast of whether two users u and v will interact with each other in the future or not [25]. Most of the already accomplished work concerned predicting links in online social media, but predicting trading interactions between users and combine several network sources in experiments is quite scarce (e.g., [16, 43]).

To contribute to this barely investigated topic the approach presented in this thesis is to predict trading interactions between users based on topological and homophilic network features of three different network sources. The data of the online social network were extracted from the public social network feeds of Second Life users. The applicable amount of information included more than $152,000$ users and more than $2,000,000$ overall feed interactions. The location-based social network information was gathered

from the in-world of Second Life by scripted avatars over a twelve-month period. They collected information about nearly 123,000 unique avatars on more than 80,000 Second Life events. The data for the trading network originate from the trading platform of Second Life, where more than 26,000 users suitable for the experiments with overall nearly 70,000 products could be found. After bringing the three different networks on a common basis to make the experiment results comparable by picking out only users appearing in all of the networks, there were 3,141 users left for the trading prediction experiments, whereof 914 were sellers and 2,776 were buyers.

For the network topological features the structures of the networks or more precisely the immediate neighbors of the involved users were relevant. For the homophilic features between user pairs in a network other characteristics were crucial. The basis for homophilic features of the online social network were the groups, interests and recorded and favored regions a user is able to declare in Second Life or the social interactions via the Second Life feeds of the users. The events the users' avatars visited in-world and their categories were the measures for homophilic features for the location-based social network. Available homophilic measures for the trading network were attributes of the products users traded with, for example, product category, product price or product rating. Various measures, for example, the cosine similarity, the Jaccard's coefficient, the Adamic Adar measure or simply the overlap of two sets were used to quantify these attributes as the similarity between two users. Overall 50 features in diverse combinations were used in the experiments to achieve trading prediction results between user pairs up to 35.90% over the baseline. Therefore, logistic regression as supervised learning method and a collaborative filtering technique as an unsupervised learning method were applied. Each experiment result also contains a table with mean values and standard errors and a correlation matrix of the involved features.

The experiments were realized in two slightly different ways. On the one hand, the origin for the prediction was a random seller and the prediction result should provide information about to what extent any random buyer will buy from the randomly chosen seller. On the other hand, the starting point was a random buyer and the prediction result should tell to what extent this buyer will buy from any random seller. In other words, the difference between these two approaches is that the direction of the trading network

changes.

1.2 Research Questions

The main topic of the thesis is trading interaction prediction. Therefore, the goal is to answer the following research questions:

Research Question 1
To what extent can trading interactions be predicted based on features obtained from trading networks? (Results: Section 6.1)

Research Question 2
To what extent can trading interactions be predicted based on features obtained from online social networks? (Results: Section 6.2)

Research Question 3
To what extent can trading interactions be predicted based on features obtained from location-based social networks? (Results: Section 6.3)

Research Question 4
Which data sources (online social network, location-based social network or trading network) and feature sets (topological or homophilic) are most suitable to predict trading interactions between users? (Results: Chapter 6.4)

Research Question 5
How do online social and location-based social network data perform compared to trading network data? (Results: Section 6.5)

Research Question 6
To what extent can the combination of trading, online social and location-based social network data increase the prediction of trading interactions? (Results: Section 6.6)

1.3 Contributions

All in all the main contributions of this work can be summarized as follows:

- Collect data of users and their activity in three different networks – an online social network, a location-based social network and a trading network
- Analyze and prepare the datasets to compute topological and homophilic features to obtain similarity measures between two users
- Show the differences and significances of the features and present two different prediction methods for trading interactions

1.4 Thesis Outline

The thesis is divided into seven chapters. Chapter 1 – this introduction – provides the problem statement and is followed by Chapter 2, which provides an overview of relevant related work in this area – networks and attributes, link prediction and virtual worlds. The dataset used in this work is described in Chapter 3 and the description of the used features is shown in Chapter 4. Chapter 5 gives a detailed description of the experimental setup with the used metrics. The results of the link prediction experiments are presented in Chapter 6. The conclusions of the master thesis is shown in Chapter 7.

2

Related Work - Link Prediction

This chapter provides an overview of relevant related work to this thesis. Already accomplished research related to different network types and their attributes right up to link prediction is summarized in this part of the work.

Liben and Kleinberg [25] defined the link prediction problem as the search to carefully predict edges that will be added to a given snapshot of a social network during a given interval. Such link predictions could be used for suggesting promising interactions between two individuals in such a social network. Also in the security and criminal investigation line of business, research already started to accent social network analysis in terms of observing terrorist networks to detect particular individuals, who are likely to work together in the future [21]. This work is concerned with the prediction of trading interactions using several user information sources similar to Guo et al. [16] as shown in Figure 2.1.

2.1 Networks

"*A person's network neighbors, taken as a whole, encompass a profoundly diverse set of relationships – they typically include family members, co-workers, friends of long duration, distant acquaintances, potentially a spouse or romantic partner, and a variety of other categories. An important and very broad issue for the analysis of on-line social networks is to use features in the available data to recognize this variation across types of relationships.*" [2]

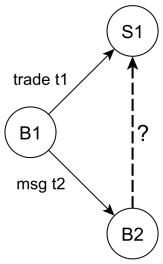


Figure 2.1: Given that B1 first purchased from S1 and then talked to B2, will B2 purchase from S1 [16]?

Within social networks, important information about users and their relations can be extracted. To interpret similarities between users it is necessary to be on familiar ground with the structure and characteristics of a network. Topological and homophilic network features are measures for such user similarities in partly large-scale network data [10, 33].

In the following subsections the features highlighted in bold letters were also applied in this thesis in a similar way.

2.1.1 Topological Features

If the structure of a network is known, network topological features could be applied to estimate the similarity between two users in the network. For their co-authorship social networks Liben and Kleinberg [25] used topological features for link prediction. As predictors they used measures as:

- ***Common neighbors***: They defined this feature easily as the number of neighbors that two users u and v have in common: $common\text{-}neighbors(u, v) = |neighbors(u) \cup neighbors(v)|$. This measure is based on the computation of Newman [28].
- ***Jaccard's coefficient***: They defined this commonly used measure for the similarity between two users u and v as $jaccard's\text{-}coefficient(u, v) = \frac{|neighbors(u) \cup neighbors(v)|}{|neighbors(u) \cap neighbors(v)|}$, as proposed by Salton and McGill [31].
- ***Adamic Adar***: Adamic and Adar [1] proposed this neighbors related measure regarding the node degree of the common neighbors. Liben and Kleinberg [25] formally defined this feature as $Adamic/Adar(u, v) = \sum_{z \in neighbors(u) \cap neighbors(v)} \frac{1}{log(|neighbors(z)|)}$.

- ***Peferential attachment***: They defined this measure, where the similarity of two users u and v is correlated with the number of neighbors of u and v, proposed by Barabasi and Albert [3], Newman [28] and Barabasi et al. 2002 [4], as $\textit{pref-att}(u,v) = |\textit{neighbors}(u)| \cdot |\textit{neighbors}(v)|$.

More detailed topological feature measures were used by Steurer and Trattner [33]. Because they partly used a directed network for their experiments in terms of predicting partnership, they distinguished between outgoing and incoming network topological features. Futures as common neighbors, total neighbors, Jaccard's coefficient and preferential attachment were each split into an outgoing and an incoming feature. Furthermore, they applied the reciprocity of user communication, Adamic Adar and the neighborhood overlap.

Additionally to common neighbors, total neighbors, Jaccard's coefficient and preferential attachment score, Fire et al. [14] defined the following topological features:

- *Transitive friends*: The size of the intersection between the outgoing neighbors of a user u and the incoming neighbors of a user v. It is formally defined as $\textit{trans-friends}(u,v) = |\textit{neighbors}_{out}(u) \cap \textit{neighbors}_{in}(v)|$.

- *Katz measure*: Katz [19] proposed a path oriented measure that defines the strength of a connection between two users u and v depending on the different paths between these users. Shorter paths result in a stronger connection. Fire et al. [14] defined it as $\textit{Katz}(u,v) = \sum_{l_{\min}=1}^{l_{\max}=\infty} \beta^l |\textit{path}^l_{u,v}|$, where $|\textit{path}^l_{u,v}|$ represents the number of paths between the users u and v and the length of each path is denoted with l. Because of the cubic complexity, which made it unpossible for them to use this feature in a large social network, they used the next feature instead.

- *Friends measure*: The friends measure they defined concerns the connections between the neighborhoods of two users u and v. They defined this features as $\textit{F-measure}(u,v) = \sum_{x \in \textit{neighbors}(u)} \sum_{y \in \textit{neighbors}(v)} \delta(x,y)$, where $\delta(x,y)$ is 1 if $x = y$ or if there is an edge between x and y and 0 otherwise.

- ***Opposite direction friends***: For directed graphs this feature indicates if reciprocal connections exist between two users u and v. They defined it as $\textit{Opp-dir-friends}(u,v) = \begin{cases} 1 & \text{if a link from } v \text{ to } u \text{ exists} \\ 0 & \text{otherwise} \end{cases}$.
- *Same community*: The value of this feature describes if two users u and v are in the same community of the network created by the Louvain method, proposed by Blondel et al. [7]. It is formally defined as $\textit{same-comm}(u,v) = \begin{cases} 1 & \text{if } u \text{ and } v \text{ are in the same community} \\ 0 & \text{otherwise} \end{cases}$.
- *Shortest path*: The value of this feature represents the shortest path length between two users u and v. They formally denoted it as $\textit{Shortest-path}(u,v)$.

Furthermore, they defined some topological features depending on subgraphs of the network.

2.1.2 Homophilic Features

Thelwall [38] described homophily as the tendency for friend- or relationships to occur between individuals. Generally, homophily is the principle that an interaction between people rather occurs if they are similar than between dissimilar people. The target of homophily is to perceive and localize the behavioral, cultural, genetic or material information that flows through networks. Homophily structures the edges of a network of every type or relationship, which could be marriage, friendship, information transfer, work advice or other types of relationship. For the personal environment common homophilic attributes are age, religion, education, occupation and gender. However, homophilic attributes are very crucial for the user behavior, the information users receive and the attitudes they form [26].

Steurer and Trattner [33] used attributes as groups, interests, user interactions, events and regions for the computation of homophilic features for their experiments for predicting partnership in social networks. For the different attributes they computed measures – also used in this thesis – such as:

- ***Common***: The number of items of an attribute two users have in common.

- ***Total***: The number of total items of an attribute of two users.
- ***Jaccard's coefficient***: They applied the Jaccard's coefficient measure on several attributes by calculating the number of common items divided by the number of total items of an attribute for two users.
- ***Cosine similarity***: For some attributes they computed the cosine similarity.

Although Fire et al. [13] used only topological features in their link prediction experiments, they demonstrated that their models surprisingly achieved considerable results. Their goal was to predict hidden links in social network structures which they tried to achieve with machine learning methods applied on several social network datasets such as Academia, TheMarker, Flickr, Youtube and Facebook. To get topological network features, the network structure has to be given. Otherwise link prediction could be applied on homophilic features, which could also be a good measure for the similarity between the users in a network, shown by Thelwall [38]. They attained highly significant indications of homophily for measures as ethnicity, age, religion, sexual orientation, country or marital status for their experiments with a MySpace dataset. What Cranshaw et al. [11] did in their work, was to combine location-based data with online social network data. They used the location-sharing Facebook application called Locaccino and tried to predict the links in the online social network. Steurer and Trattner [33] also combined online social network data with location-based social network data in their partnership prediction experiments.

2.2 Learning Methods

In the literature typically two different learning methods are commonly used to predict links between users in a network. The same approaches are also used in the thesis:

2.2.1 Supervised Learning

Hasan et al. [17] considered a social network with interactions as edges representing the coauthoring of research articles. Each article included author

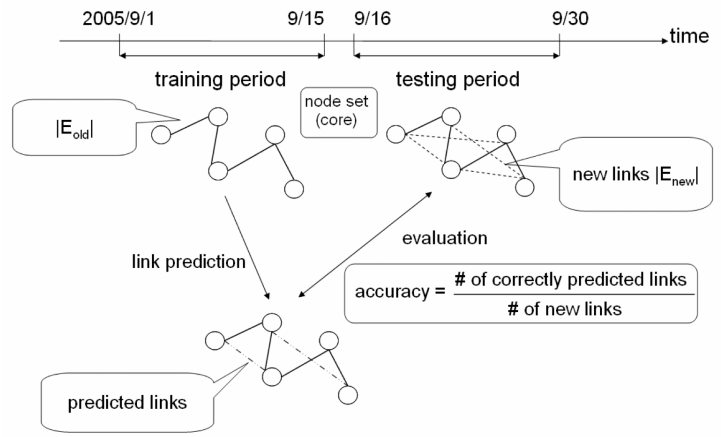


Figure 2.2: Overall procedures for their experiments [27].

information and publication year, at least. To make a link prediction, they first split the set of publication years into two non-overlapping sub-ranges as training and test set. Their classification dataset consisted of author pairs that already existed in the trainings set, but did not publish any papers together in this period. To become a positive example for their experiment, those author pairs had to publish at least one paper in the test set period, otherwise they represented a negative example. Each positive example of author pairs established a link between them, which did not exist for the period of the training set. Consequently, they had a binary classification problem that was solved by supervised learning.

Murata and Moriyasu [27] showed a graphical overview of the basic procedure of a link prediction experiment resulting in a value for the performance of the used experimental setup (see Figure 2.2).

Steurer and Trattner [33, 34, 35], Steurer et al. [36] and Trattner et al. [39] used the following learning algorithms for the experiments in their works:

- *Decision Trees (C4.5) - J48* [9, 13]
- *Logistic Regression* [9, 18, 23, 30, 43]
- *Random Forest* [14, 18]
- *Support Vector Machines (SVM)* [13, 17, 18, 43]

Most of the time the best results were achieved with logistic regression.

Table 2.1: Area under the ROC curve (AUC) for predicting partnership with different feature sets and learning algorithms. The best algorithm for each feature set is highlighted in bold letters [33].

Feature Sets		J.48	Logistic Regression	SVM
Online Network	Topological	**0.823**	0.743	0.659
	Homophilic	0.775	**0.817**	0.720
	Combined	0.860	**0.878**	0.771
Location-Based Network	Topological	0.745	**0.772**	0.657
	Homophilic	0.852	**0.902**	0.818
	Combined	0.845	**0.905**	0.829
Combined		0.881	**0.933**	0.859

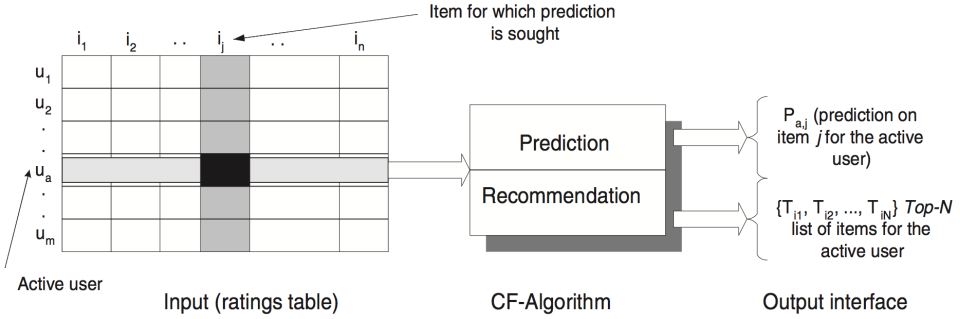


Figure 2.3: The fundamental process of collaborative filtering [32].

For their partnership prediction Steurer and Trattner [33] reached AUC values up to 43.30% over the baseline (see Table 2.1).

2.2.2 Unsupervised Learning

To predict links with unsupervised learning, ranked lists of the potential links have to be established using the available network quantities. Then the k top-ranked potential links have to be classified as new links and the rest as missing links with k as the number of expected new links [42]. It is useful to compare the results of the supervised learning approach with an unsupervised learning method as collaborative filtering [20, 29, 32] to substantiate the achieved prediction performance or indicate additional support for the used features [6]. Figure 2.3 shows the fundamental process of collaborative filtering with the users u on the left side of the matrix and the items i on the top. For the prediction problem in this thesis the items are also users.

Steurer and Trattner [33] applied this method for their partnership pre-

diction. First, they reduced the link prediction problem to a binary classification problem by selecting the same amount of positive and negative edges and applied different supervised learning algorithms to a training set and verified this in a test set. Second, they substantiated their results with a collaborative filtering technique by establishing a ranked list with all acquaintances for every user according to the different features and computing the mean value of the success rates of finding the positive edges in these ranked lists.

3 Dataset

In this chapter the used datasets for the realized experiments are described. As depicted in the introduction, all experiments were conducted in the context of the virtual world Second Life. There were several reasons for this decision. To answer the research questions it was necessary to possess information about different sources of data. Data from an online social network such as Facebook[1] or Google+[2], from a location-based social network as Foursquare[3] and from a trading network as eBay[4] were needed. The problem with the even mentioned resources was that they restrict the verbose crawling of their user profiles, but apart from this awareness, most of the users share their profiles only with their friends and prohibit the access by others. Another important point was the possibly sparse amount of overlapping data of the several networks, because of the different participants. [33, 34, 35, 36]

3.1 Virtual World - Second Life

"*Second Life (SL) is a virtual world where people interact and socialize through virtual avatars. Avatars behave similarly to their human counterparts in real life and naturally define a social network.*" [41]

The basic principle of Second Life is that avatars explore the virtual world,

[1] `http://facebook.com/`
[2] `http://plus.google.com/`
[3] `http://foursquare.com/`
[4] `http://ebay.com/`

meet other avatars and communicate, play or trade with them. In this virtual world there are not solely human-controlled avatars, but also bots – automated avatars, which may be difficult to differentiate from human-controlled avatars and can damage the user experience in Second Life. Such bots can be controlled with automated scripts in a variety of ways as welcoming other avatars to a region or spying on user behavior in Second Life. Varvello and Voelker [41] denoted the Second Life social network as small-world network and much more similar to a natural network in comparison with popular online social networks. Crucial for this observation is the establishing of social relationships between users in Second Life, which expect an active interaction between the involved users. By contrast, relationships in online social networks often signify only the acceptance of a friendship request without existing interactions as text messages between the users.

The virtual world of Second Life consists of regions – 256x256 meters in size and independent from each other – that the users' avatars can enter. The owner of such a region, who can either be Linden Lab[5] – the developer of Second Life – or an individual or a company, has full control over the land and is able to, for example, limit the access to the region to a selected set of avatars or define a specific policy for object creation. The usage or purpose of a region depends on the objects it contains. There are regions called *Sandbox* in Second Life, for instance, that enable avatars to build and test new objects or scripts. The Linden Lab servers that host regions are called *Simulators*. They manage the state of their regions, provide information about the objects and land features and handle the chat between the currently connected avatars [40].

To participate in this virtual world users have to register online with Second Life on the Second Life website[6], first and then download the *Second Life viewer*[7] – a client software of Linden Lab – as shown in Figure 3.1 or any available third party client, e.g. *libopenmetaverse*[8]. After a successful login the users' avatars are able to walk, run, fly or even teleport around the regions of the in-world of Second Life or, for instance, chat with others. The Second Life viewer provides the user with a limited field of view of a 35 meter radius. Furthermore, the user has the possibility to face a visual

[5] `http://lindenlab.com/`
[6] `http://secondlife.com/`
[7] `http://secondlife.com/support/downloads/`
[8] `http://lib.openmetaverse.org/`

Figure 3.1: A screenshot of the Second Life viewer.

Figure 3.2: An example of the Second Life map with a region and some connected avatars.

overview of all regions in Second Life in terms of a map. Green dots on the maps symbolize the coordinates of the currently connected avatars in each region, as the example in Figure 3.2 shows [40].

As online social network source the *My Second Life*[9] network was used, the location-based data were monitored in-world in Second Life and the *Second Life Marketplace*[10] was used as trading network in this thesis. Consequently, the users of these three different network sources have in common that they are all Second Life users. Although they do not interact with their real

[9] `http://my.secondlife.com/`
[10] `http://marketplace.secondlife.com/`

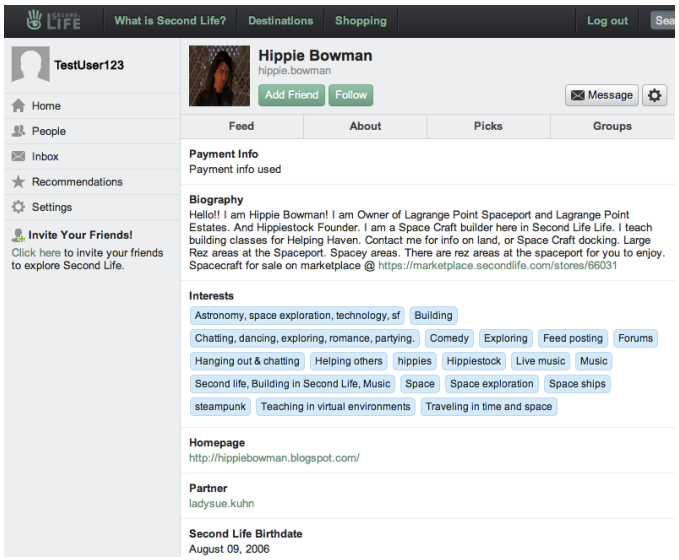


Figure 3.3: A Second Life user's about tab containing a.o. the biography, the interests and the Second Life birthdate.

life names, but with the names of their avatars in a virtual world, La and Michiardi [22] and Varvello et al. [40] have shown that the avatars' behavior tend to be similar to humans' behavior.

3.2 Online Social Network Data

Similar to the real world users in the virtual world of Second Life are also able to establish social links through an online social networking platform called My Second Life. It was introduced by Linden Lab in 2007 and can be compared with other online social networks as Facebook or Google+. This platform gives Second Life users the opportunity to present personal information on their user profiles or interact with other users on the so-called *Feed*, which can be compared with the *Wall* in Facebook. Apart from such information as belongs to the Second Life avatar such as interests, the day of birth in Second Life or the biography (see Figure 3.3), users are able to join groups (see Figure 3.4) or show their favorite in-world regions on their profiles. It is also possible to share text messages or pictures with others on the feed. Furthermore, these postings can be commented or loved (see Figure 3.5). A *Love* in Second Life is similar to a *Like* in Facebook or a *Plus* in Google+. A considerable difference to Facebook exists concerning

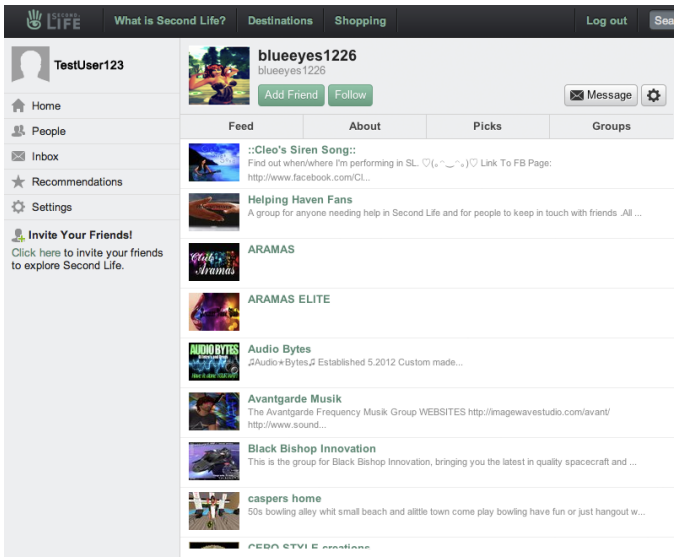


Figure 3.4: List of joined groups of a user in Second Life.

friendship relations. Such relation type does not exist in My Second life, and this is why there is no need to send a friend request that the other one has to confirm as it is the case in Facebook. So the restriction that only friends are allowed to post on the feed of a user is not given, instead every user is able to post on each users' feed [33].

At the end of March 2013 the Second Life profiles of users who had not changed their privacy settings to private were crawled by a web crawler – implemented in the widely used, high-level programming language *Python*[11]. A list of user names was extracted from the location-based dataset (see Section 3.3) and iteratively extended by further users who interacted on the feed with the users from the list. For each user the user's *interests*, the joined *groups* and the *feed interactions* with others were ascertained. Two different sources of Second Life regions were also part of the collected information for each user. It is possible to record in-world snapshots of Second Life regions in terms of pictures and share them on the feed to show others where users have actually been at a particular time (see Figure 3.5). These so-called *recorded regions* were gathered for each user and thus count as personal user information. Besides the interests, groups, biography etc., the profiles in Second Life provide an area to state preferred in-world locations – the second source of locations and so-called *favored regions* (see Figure 3.6).

[11] `http://python.org/`

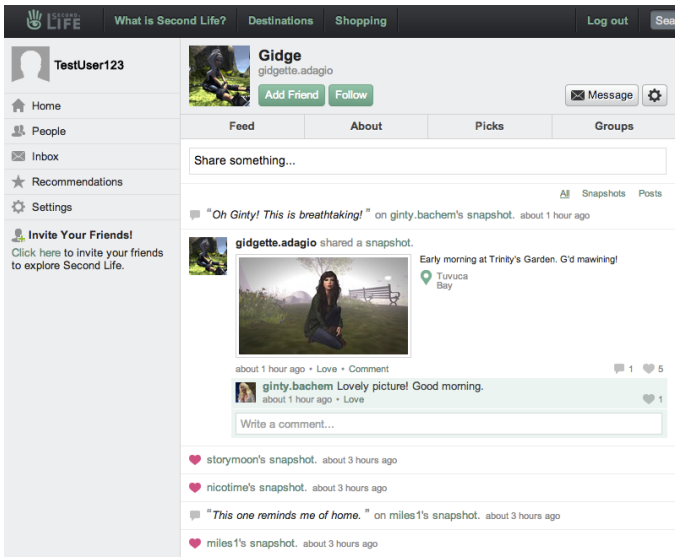


Figure 3.5: Example of a Second Life feed where users can post text messages or pictures or comment or love each other's postings. There is also a recorded Region on that feed called "Tuvuca Bay".

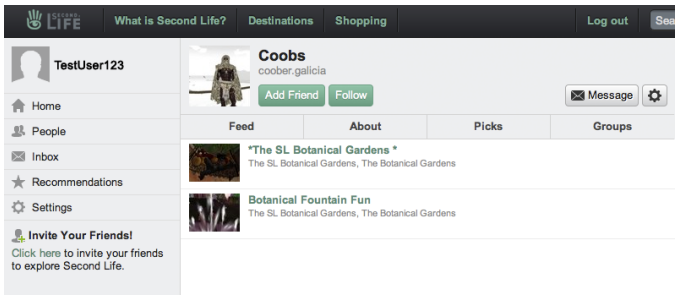


Figure 3.6: The picks tab on a Second Life user's profile showing the self-defined favored regions.

The online social network was constructed on the basis of the feed interactions between the users, which were an indicator for an acquaintance between them. If the number of interactions was zero, there was no link between them. Users with numbers of interactions greater or equal one were provided with an edge between them in the network. Eventually, this directed online social network was denoted as $G_O = \langle V_O, E_O \rangle$, where V_O was the set of users with interactions on their feed. If a user $u \in V_O$ communicated with a user $v \in V_O$ by posting a text message on v's feed or commenting or loving a posting on v's feed, the edge between them was formally defined

as $e = (u, v) \in E_O$.

First, this procedure reached a result of 169,035 users with 587,090 postings, 459,734 comments and 1,631,568 loves, which gave a number of total interactions of 3,175,304. Each user defined ≈1.5 interests and joined ≈12.4 groups on average. The total number of unique interests was 67,644 and the total number of unique groups was 214,363. Overall, 13,583 unique recorded regions and 25,311 unique favored regions were found. On average, each user posted ≈2.9 recorded regions on the feed and stated ≈4.8 favored regions in the Second Life profile.

Due to the fact that the master thesis is about predicting trading interactions, self connections in the network have been removed, because seller and buyer are not the same person in a trading relation. In this way, the dataset for the online social network of Second Life slightly decreased. Now there were 152,509 users with 226,668 postings, 348,106 comments and 1,494,044 loves, which gave a number of total interactions of 2,068,818. Probably, the number of loves remained nearly stable, because the loves for postings mostly apply other users' postings and not one's own. Furthermore, the total number of unique interests shrank to 62,170 and the number of unique groups to 204,769. The average of the number of interests defined by each user stayed the same with ≈1.5 and the joined groups per user slightly changed to ≈12.4 on average. The number of unique recorded and favored regions almost stayed the same, but the average of the number of recorded regions per user changed to ≈3.1 and of stated favored regions to ≈2.2.

Table 3.1 gives an overview of the numbers of the online social dataset and Figure 3.7 displays a visualization of the structure of the online social network showing the high amount of connected components. For simplification only a randomly chosen subset of the network of about 1,000 edges was illustrated. Among others, the basic network statistics for this online social network are displayed in Table 3.4 and the power-law distributions for sales and purchases are shown in Figure 3.16.

Table 3.1: Statistics of the online social network dataset.

Online Social Network G_O	
Users	152,509
Edges	270,567
Postings (Text Messages / Pictures)	226,668
Comments	348,106
Loves	1,494,044
Overall Interactions	2,068,818
Average Interactions per User	≈14
Group Joins	1,869,281
Unique Groups	204,769
Users with Group Join(s)	114,205
Stated Interests	227,596
Unique Interests	62,170
Users who stated Interest(s)	36,610
Recorded Regions Postings	466,930
Unique Recorded Regions	13,251
Users with Recorded Postings	36,430
Stated Favored Regions	337,732
Unique Favored Regions	22,742
Users who stated Favored Region(s)	76,093

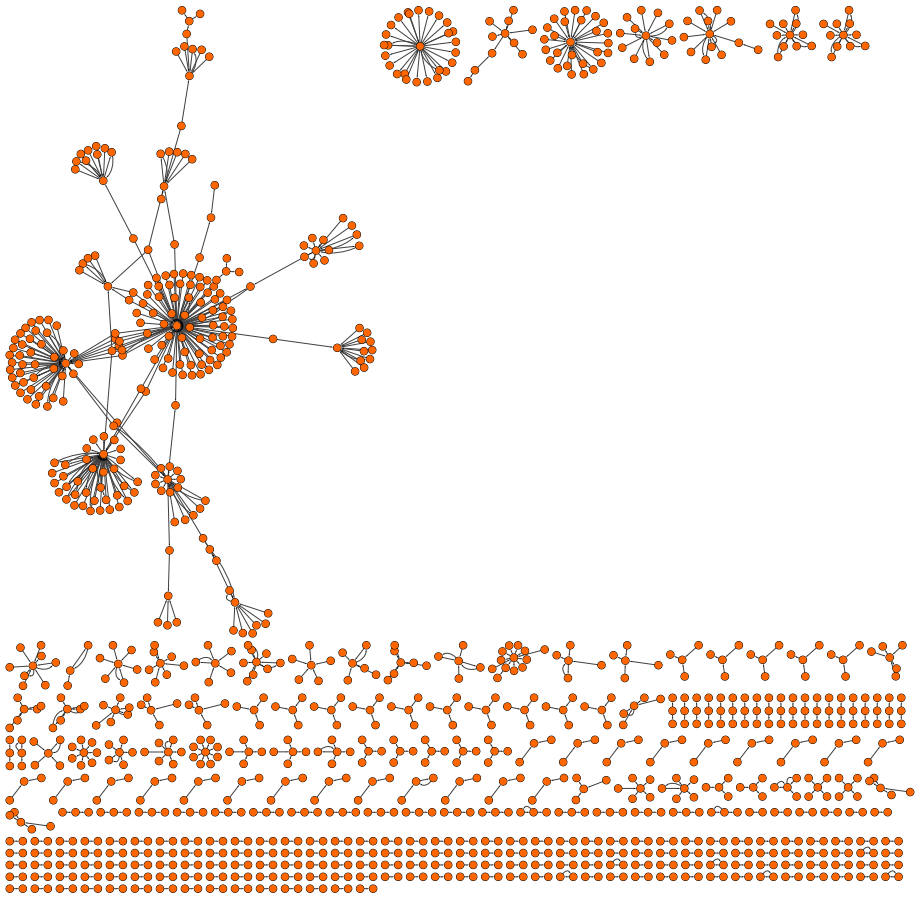


Figure 3.7: Network structure of a subset ($\approx 1,000$ edges) of the online social network with a high amount of connected components.

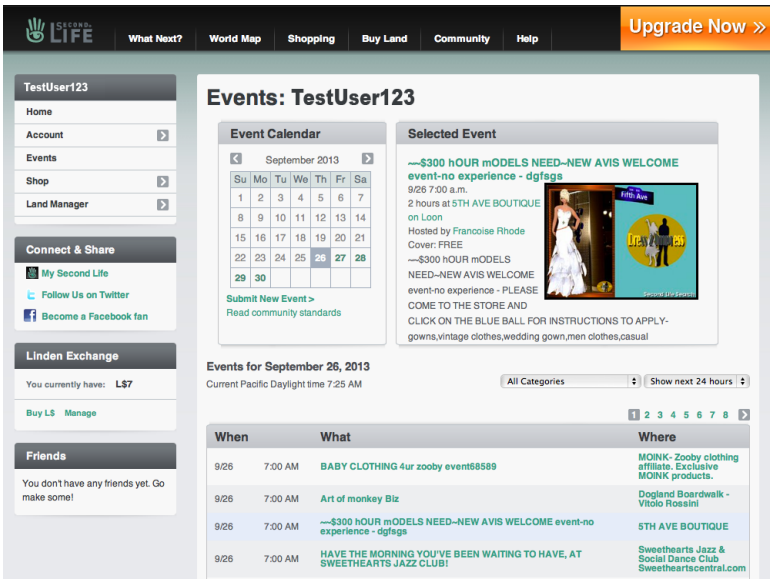


Figure 3.8: The Second Life events calendar where users can choose the date, category and time interval to get a list of corresponding events.

3.3 Location-Based Social Network Data

The location-based dataset used in this master thesis was extracted from the in-world of Second Life by scripted robots visiting much-frequented locations.

Similarly to the real world, also in Second Life, most of the interactions between people during the day happen on events. Users have the possibility to create events and advertise them to the public in the Second Life event calendar (see Figure 3.8). Each event contains among others name, description, start and end time as well as location and is provided with one of eleven predefined categories, e.g. *Nightlife/Entertainment* or *Arts and Culture*. These events were useful, because they presumably afforded a better user frequency than other places in the huge world of Second Life. The average distribution of the events for different event categories over one day with two peaks at 12:00 pm and 6:00 pm is shown in Figure 3.9. The x-axis represents the timeline for the day and the y-axis number of events.

From March 2012, a simple web crawler parsed the events website for the duration of twelve month and stored the events information in a database. Simultaneously, adopted libopenmetaverse clients – written in the programming language *C#*[12] – operated in-world as user data collecting bots. They

[12] `http://msdn.microsoft.com/en-us/vstudio/hh341490/`

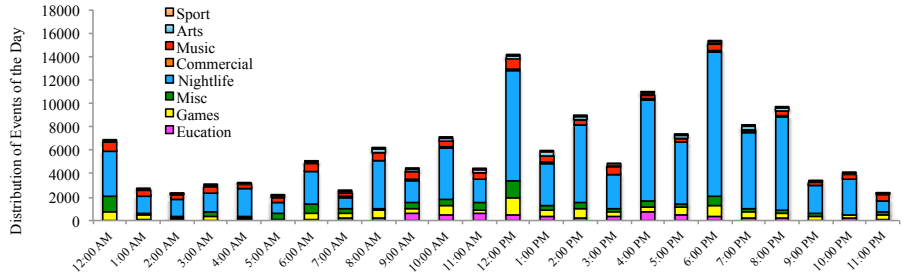


Figure 3.9: Distribution of the events for different event categories over one day [37].

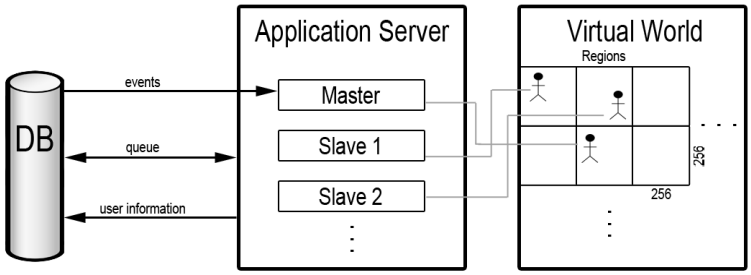


Figure 3.10: Master-Slave architecture with the database, the application server including the bots and the virtual world of Second Life.

had the ability to automatically move around, jump more exactly from one location to another, and read out information about surrounding users. With a master-slave architecture about ten instances were transmitted in-world at once. The master robot had the task to prepare a queue of currently happening events in the local *MySQL*[13] database and the slave robots visited these events in 15 minute intervals and stored information about the attendees, e.g. user name, location and time (see Figure 3.10).

This information which was collected over the period of one year, was the basis for the location-based social network. Overall, nearly 19 million data entries with 410,619 different users in 4,146 different locations were observed. To generate a network with this huge amount of data without going beyond the scope of the density of the network, it was necessary to create a link between two users only then, if they had met each other in the same location at the same time on at least two different days. This rule reduced the number of edges in this network many times over to 1,414,389

[13] `http://mysql.com/`

Table 3.2: Statistics of the location-based social network dataset.

Location-Based Social Network G_L	
Users	122,936
Edges	1,414,389
Events Entries	1,966,206
Unique Events	81,671
Event Categories	11
Average Events per User	≈16
Event Regions Entries	16,375,540
Unique Event Regions	3,972
Average Entries per User	≈133

and the number of nodes to 122,936. The total number of monitored events for all users was 1,966,206 with 81,671 unique events and eleven different event categories. That gives an average of ≈16 events per user. There were 16,375,540 event regions entries registered with 3,972 unique regions, which means that on the average, each user was found ≈133 times by the robots.

Table 3.2 gives an overview of the numbers of the location-based dataset and Figure 3.11 displays a visualization of the structure of the location-based social network. In comparison to the online social network (Figure 3.7) the density is higher and so the number of connected components is much smaller. For simplification only a randomly chosen subset of the network of about 1,000 edges was illustrated. Among others, the basic network statistics for this undirected location-based social network are displayed in Table 3.4 and the power-law distributions for sales and purchases are shown in Figure 3.16. It was formally defined as $G_L = \langle V_L, E_L \rangle$, where V_L was the set of users and $e = (u, v) \in E_L$ the link between two users $u \in V_L$ and $v \in V_L$, if they were observed together in the same place at the same time on at least two different days.

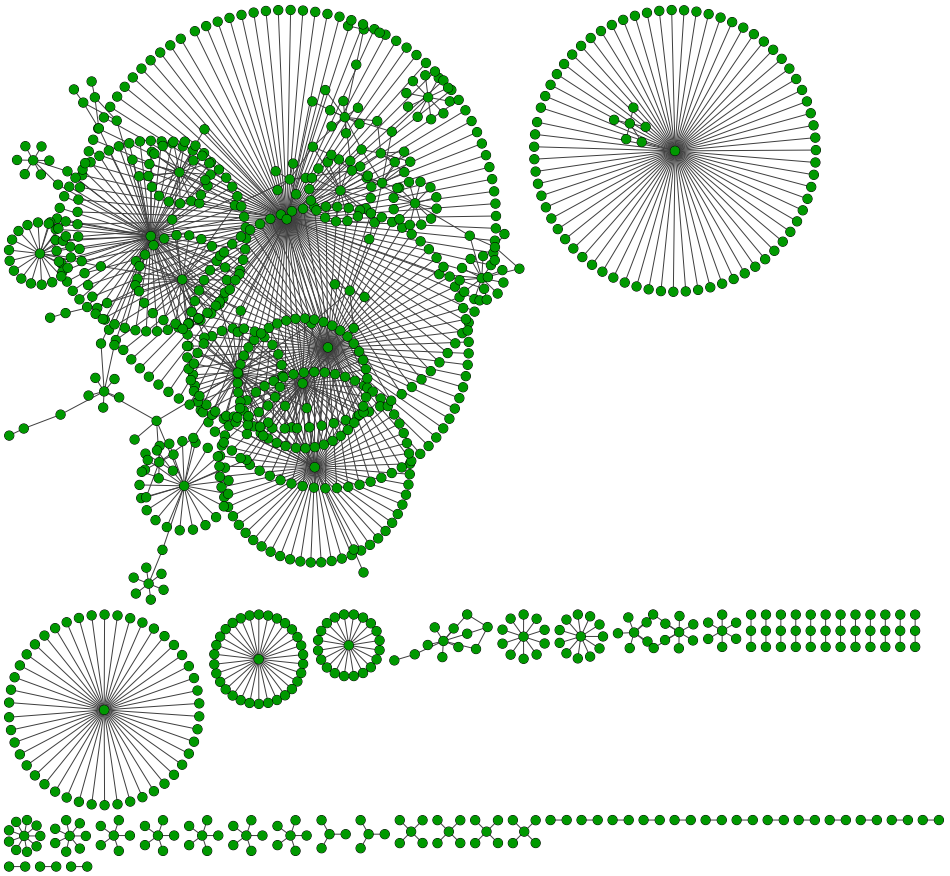


Figure 3.11: Network structure of a subset ($\approx 1,000$ edges) of the location-based social network with a high density.

3.4 Trading Network Data

Second Life provides an online trading platform called Second Life Marketplace where Second Life users are able to trade with virtual goods (see Figure 3.12). Basically, the credit balance of the users is empty after registering to Second Life, but similarly to the real life it is necessary to also have a currency in virtual worlds. *Linden dollar* (L$) is the Second Life's currency with a defined exchange rate between it and the U.S. Dollar (US$). To convert real world money to virtual world money and vice versa there are several online exchange platforms, e.g. VirWoX[14]. Virtual money in Second Life is useful to buy or sell objects in-world or via the Marketplace.

[14] `http://virwox.com/`

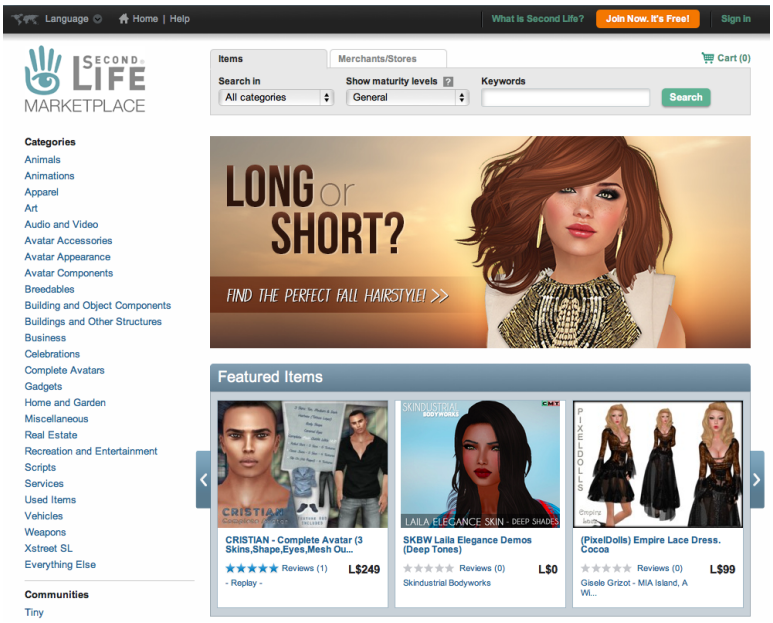


Figure 3.12: A screenshot of the main page of the Second Life Marketplace web site.

This online trading platform gives sellers an opportunity to provide their offered products with pictures, a descriptions and a price and to classify them into different product categories as shown in Figure 3.13. Similarly to online shopping platforms such as eBay, a user can be a seller, a buyer or both. It has to be said that the sale of products requires the creation of a store in the Marketplace. So every seller in the Second Life Marketplace owns a separate store which contains all offered products (see Figure 3.14). Only if a purchase is done via the Marketplace, the buyer can write a public review about the bought product or just rate the product from one to five stars. As a consequence, every stated review in the whole marketplace ensures the purchase of the product between the seller and the reviewer. Linking all sellers with their buyers based on the product reviews was the basic idea for the trading network for the experiments in this thesis.

To collect this kind of information all store sites of the Second Life Marketplace were gathered with a Python web crawler. Because of the fact that each store has its unique id, it was easy to catch all existing stores by instructing the crawler simply to iterate the store id inside the web addresses of the stores from zero until no more stores could be found. So all Second Life store's web sites were crawled and stored in the local database. To get

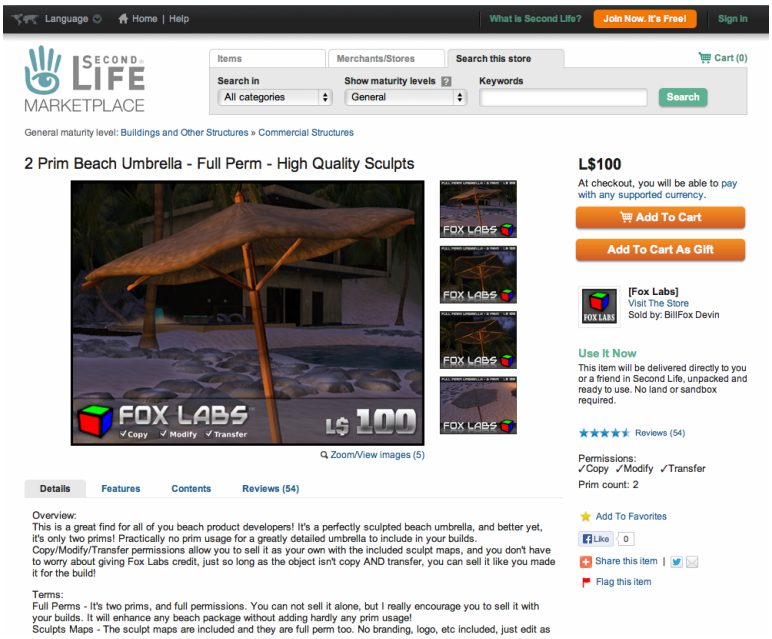


Figure 3.13: A screenshot of the detailed site of a Second Life product within the Second Life Marketplace.

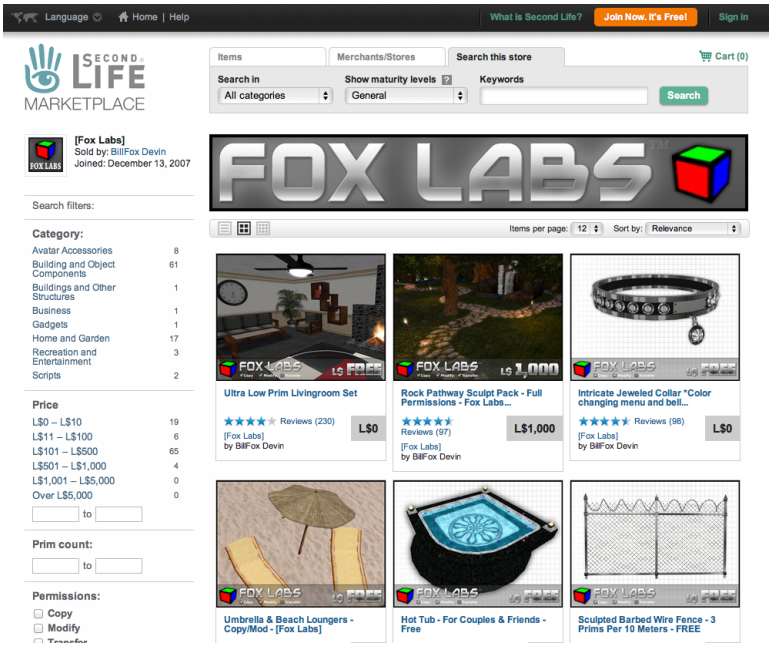


Figure 3.14: A screenshot of the overview of a Second Life store within the Second Life Marketplace.

detailed product information of all stores it was necessary to additionally crawl each product site separately as well, because some of the information such as for instance when the product reviewers (buyers) were not declared in the store overview site, but in the detailed product site. Thus the web crawler collected the information of all associated products. Therefore, all sellers, their appropriated buyers and the purchased products containing the price, category and the ratings were known. With this procedure the crawler detected 131,087 stores/sellers, whereof 36,330 had at least one product in supply and 17,914 sold at least one product. Overall 1,725,449 products in 22 different categories, e.g. *Avatar Accessories* or *Vehicles*, were found, from which 120,762 were purchased at least once. The total number of noticed purchases was 268,852 with 77,645 different buyers. Due to the fact that a seller can also be a buyer and a buyer can also be a seller, 8,259 users acted as both seller and buyer. The total number of involved users was 87,300.

The third network in this thesis was generated by linking all sellers with their buyers. For the experiments it was essential to avoid *overfitting*. Overfitting occurs if the data fits too well and clarifies the random variation and also the significant relationships in it [8]. Both sellers and buyers with a total number of purchases below three were excluded from the network to avoid overfitting, because the result for predicting a trading interaction between two users, having insufficient purchase information about them, would not have been valid. For example, the prediction probability for trading interactions based on features such as *Cosine Similarity of Product Categories* (see Section 4.3) would have been too high, because for user pairs where both the seller and the buyer have, for example, only one purchase to report, the cosine similarity for the product categories could only be 1, if their purchased products were of the same category, or 0 otherwise. This example shows that without a threshold the amount of purchase information for such user pairs would have been insufficient for a valid prediction. Certainly the size of data shrank with the purchase threshold $t = 3$ – the condition that each user must have made at least three purchases – down to 26,097 users, from which 8,365 were sellers, 20,819 were buyers and 3,087 were both sellers and buyers. The number of involved products subsided down to 69,982 and now the number of total purchases was 140,475.

Table 3.3 gives an overview about the numbers of the trading dataset and Figure 3.15 displays a visualization of the structure of the directed trading

Table 3.3: Statistics of the trading network dataset with threshold $t = 1$ and $t = 3$.

	Trading Network G_T	
	$t = 1$	$t = 3$
Users	87,300	26,097
Edges	219,889	105,778
Stores/Sellers	17,914	8,365
Buyers	77,645	20,819
Sellers + Buyers	8,259	3,087
Product Categories	22	22
Products	120,762	69,982
Average Products per Seller	≈7	≈8
Purchases	268,852	140,475
Average Purchases per Seller	≈15	≈17
Average Purchases per Buyer	≈3	≈7

Table 3.4: Basic metrics of the online social network, the location-based social network and the trading network.

	Online G_O	Location-Based G_L	Trading G_T
Type	directed	undirected	directed
Nodes	152,509	122,936	26,097
Edges	270,567	1,414,389	105,778
Degree	3.55	23.01	8.11
Connected Components	13,115	719	77
Largest Connected Component	77.69%	99.01%	99.26%
Sellers	6,627	4,319	8,365
Buyers	22,718	20,120	20,819
Sellers + Buyers	3,777	2,458	3,087

network with a much lower average degree compared to the location-based social network (Figure 3.11). For simplification only a randomly chosen subset of the network of about 1,000 edges was illustrated. Among others, the basic network statistics for this directed trading network are displayed in Table 3.4 and the power-law distributions for sales and purchases are shown in Figure 3.16. It was formally defined as $G_T = \langle V_T, E_T \rangle$, where V_T was the set of users and $e = (u, v) \in E_T$ the link between two users $u \in V_T$ and $v \in V_T$, if they v is a buyer of u.

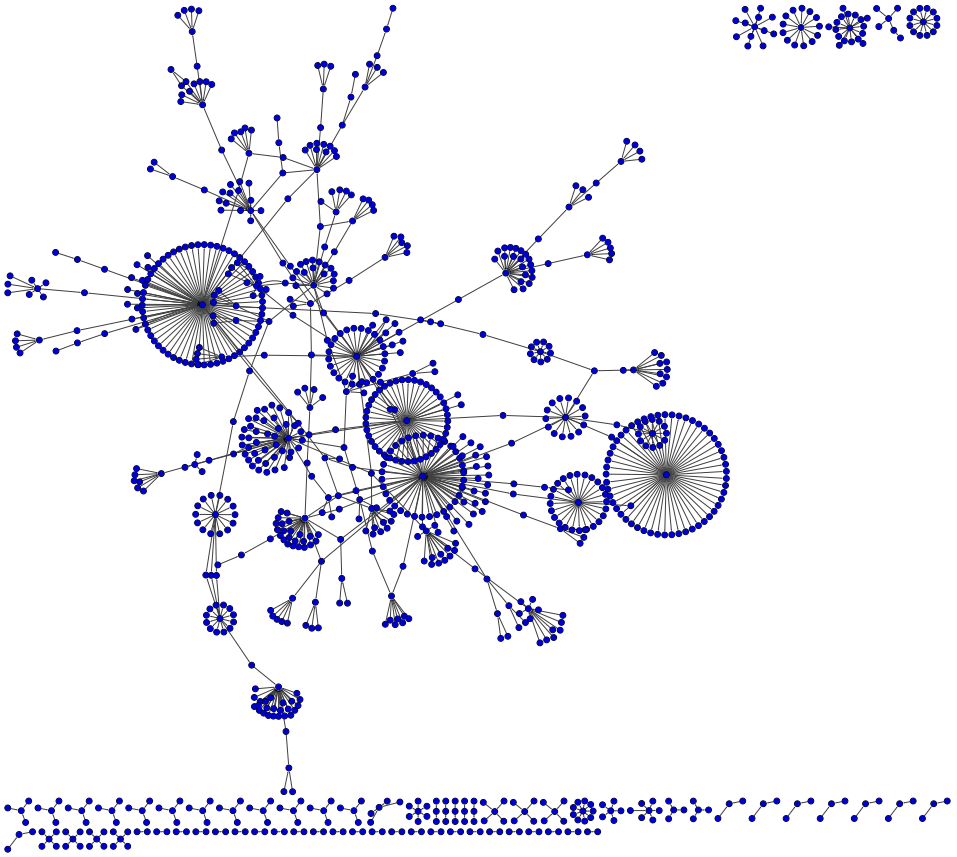
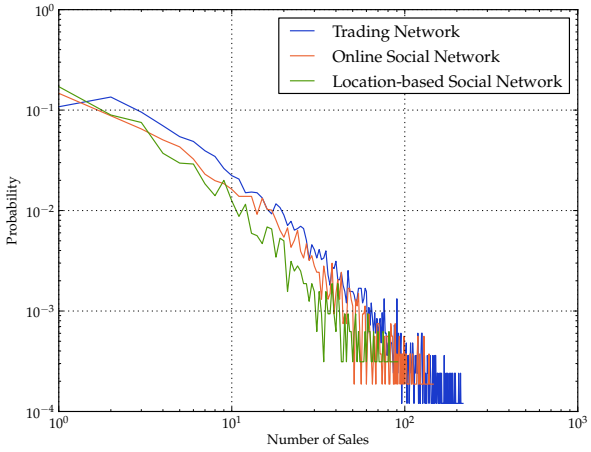
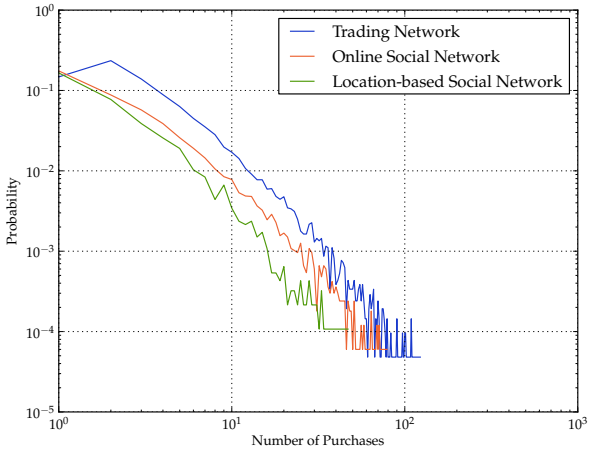


Figure 3.15: Network structure of a subset ($\approx 1,000$ edges) of the trading network.



(a) Distributions of the number of sales of each seller in all three networks.



(b) Distributions of the number of purchases of each buyer in all three networks.

Figure 3.16: Power-law distributions [12] of the number of sales and purchases for all three networks.

4 Feature Description

This chapter elaborates in detail the features used and extracted for the experiments in this thesis. As mentioned in Chapter 2, different characteristics can be extracted from a network and used for further similarity computations between users. Each calculated feature is a measure for the similarity of two users' attributes, e.g. the number of groups they have in common or the number of their common neighbors in a network. Each of the three networks has both network topological and homophilic features. Topological features try to capture similarities between users by only taking the network structure into account. For example, to get the total number of neighbors of two users all their surrounding users in the network have to be included in the computation. In contrast, homophilic features only need the particular attribute of the two involved users. For example, to get the number of common interests, the sets of interests of both users simply have to be compared [30, 34].

4.1 Online Social Network Features

This section provides a description of the features derived from the online social network dataset.

Topological Features

The neighbors of a user u in this directed network were defined with respect to the direction of the communication between them. A neighbor v that received messages from a user u is called *outgoing neighbor* and a neighbor v

that sent messages to a user u is called *incoming neighbor* [33]. The definition of outgoing neighbors of a user $u \in V_O$ was denoted as $\Theta^+(u) = \{v \mid (u,v) \in E_O\}$ and incoming neighbors as $\Theta^-(u) = \{v \mid (v,u) \in E_O\}$. So the whole set of neighbors of u could be formally computed as $\Theta(u) = \Theta^+(u) \cup \Theta^-(u)$.

- *Common Outgoing Neighbors*: The number of neighbors that two users u and v have in common related to the outgoing communication of them was defined as $O^+_{CN}(u,v) = |\Theta^+(u) \cap \Theta^+(v)|$. For example, a user $w \in O^+_{CN}(u,v)$ is a common outgoing neighbor of user u and v if both u and v sent one or more messages to w. This and the next feature were taken from Steurer and Trattner [33].

- *Common Incoming Neighbors*: This is the opposite of the common outgoing neighbors. The number of common incoming neighbors of two users u and v are the users who sent messages to both of them. This feature was defined as $O^-_{CN}(u,v) = |\Theta^-(u) \cap \Theta^-(v)|$. For example, a user $w \in O^-_{CN}(u,v)$ is a common incoming neighbor of user u and v if w sent one or more messages to u as well as v.

- *Total Outgoing Neighbors*: The difference to the common outgoing neighbors is that this feature shows the number of the union set of all neighbors in the outgoing direction of two users u and v. It was defined as $O^+_{TN}(u,v) = |\Theta^+(u) \cup \Theta^+(v)|$.

- *Total Incoming Neighbors*: This feature describes the number of the union set of all neighbors related to the incoming communication of two users u and v and was denoted as $O^-_{TN}(u,v) = |\Theta^-(u) \cup \Theta^-(v)|$.

- *Outgoing Jaccard's Coefficient*: The Jaccard's coefficient is the division of the common by the total neighbors of two users u and v and could be seen as a measure for exclusiveness of the relation between them [11]. It was also split into an outgoing and an incoming feature. The outgoing Jaccard's coefficient was denoted as $O^+_{JC}(u,v) = \frac{|\Theta^+(u) \cap \Theta^+(v)|}{|\Theta^+(u) \cup \Theta^+(v)|}$.

- *Incoming Jaccard's Coefficient*: This feature is the complement to the outgoing Jaccard's coefficient and was defined as the common incoming neighbors divided by the total incoming neighbors: $O^-_{JC}(u,v) = \frac{|\Theta^-(u) \cap \Theta^-(v)|}{|\Theta^-(u) \cup \Theta^-(v)|}$.

- *Preferential Attachment Out*: Here the preferential attachment score, first proposed by Barabasi and Albert [3], is presented in a slightly different way, proposed by Cheng et al. [9]. It is another popular measure to describe the correlation between the out-degree of a user u and the in-degree of a user v. The value for this feature was calculated as the product of the number of outgoing neighbors of u and the incoming neighbors of v, formally defined as $O^{+}_{PS}(u,v) = |\Theta^{+}(u)| \cdot |\Theta^{-}(v)|$.

- *Preferential Attachment In*: The difference to the preferential attachment out feature described above is that the in- and out-degree of the involved users were swapped. So the preferential attachment in feature for two users u and v was denoted as $O^{-}_{PS}(u,v) = |\Theta^{-}(u)| \cdot |\Theta^{+}(v)|$.

- *Reciprocity of User Communication*: The reciprocity of user communication in a directed network describes if a communication between two users u and v is bidirectional or in only one direction [9]. This feature was denoted as $O_R(u,v) = \begin{cases} 0 & \text{if } (u,v) \in E_O,\ (v,u) \notin E_O \\ 1 & \text{if } (u,v) \in E_O,\ (v,u) \in E_O \end{cases}$.

- *Adamic Adar*: A sophistication of the relation between two users related to their neighbors was proposed by Adamic and Adar [1]. It is a measure for the activity of the common neighbors of two users u and v in the network, because the definition regards the node degree of the common neighbors. For directed networks Cheng et al. [9] suggested a refinement of the Adamic Adar measure in which only the common incoming neighbors are considered:
$O_{AA}(u,v) = \sum_{z \in \Theta^{-}(u) \cap \Theta^{-}(v)} \frac{1}{log(|\Theta(z)^{-}|)}$.

Homophilic Features

The groups a user u can join in this social network were defined as $\Delta(u)$ and the self-defined interests of u as $\Phi(u)$. The following features were taken from Steurer et al. [36] and Steurer and Trattner [33]:

- *Common Groups*: This feature represents the number of groups two users u and v have in common: $G_C(u,v) = |\Delta(u) \cap \Delta(v)|$.

- *Total Groups*: The number of the union set of the joined groups of two users u and v was defined as $G_T(u,v) = |\Delta(u) \cup \Delta(v)|$.

- *Jaccard's Coefficient of Groups*: The Jaccard's coefficient as already mentioned can also be applied for homophilic measures such as groups, interests, regions or events. In this case the Jaccard's coefficient for groups was denoted as $G_{JC}(u, v) = \frac{|\Delta(u) \cap \Delta(v)|}{|\Delta(u) \cup \Delta(v)|}$.

- *Common Interests*: The same types of features as defined for groups were determined for the interests users are able to declare on their social feed. The common interests feature shows the number of interests two users u and v have in common: $I_C(u, v) = |\Phi(u) \cap \Phi(v)|$.

- *Total Interests*: A total feature was defined for interests, too. For two users u and v the total number of features was formally defined as $I_T(u, v) = |\Phi(u) \cup \Phi(v)|$.

- *Jaccard's Coefficient of Interests*: For the user-defined interests the Jaccard's coefficient for the proportion of common and total interests of two users u and v was denoted as $I_{JC}(u, v) = \frac{|\Phi(u) \cap \Phi(v)|}{|\Phi(u) \cup \Phi(v)|}$.

- *Interactions*: In the online social network of Second Life the users are able to share text messages with other users or comment or love such messages. The interactions from a user u to a user v were defined as $\iota(u, v)$. So this feature shows the number of all interactions from u to v and was formally defined as $OI(u, v) = |\iota(u, v)|$.

On the online social network feed of Second Life users are able to record in-world snapshots of regions in terms of pictures and share them to show their friends or followers where they have actually been at a particular time. Such regions a user u shared on the feed were denoted as $\Lambda(u)$. This information of data was used for the next features, borrowed from Steurer et al. [36] and Cranshaw et al. [11]:

- *Common Recorded Regions*: This feature is a measure for how many common recorded regions two users u and v shared on their own feed and was formally specified as $RR_C(u, v) = |\Lambda(u) \cap \Lambda(v)|$.

- *Total Recorded Regions*: The number of the union set of such recorded regions of two users u and v: $RR_T(u, v) = |\Lambda(u) \cup \Lambda(v)|$.

- *Jaccard's Coefficient of Recorded Regions*: The value of the common recorded regions divided by the value of the total recorded regions of

two users u and v is the Jaccard's coefficient measure again and was defined as $RR_{JC}(u,v) = \frac{|\Lambda(u) \cap \Lambda(v)|}{|\Lambda(u) \cup \Lambda(v)|}$.

- *Overlap of Recorded Regions*: The overlap of the sets of recorded regions of two users u and v differs from the Jaccard's coefficient in terms of the division by the sum of u's and v's regions. So this feature was defined as $RR_O(u,v) = \frac{|\Lambda(u) \cap \Lambda(v)|}{|\Lambda(u)| + |\Lambda(v)|}$.

Apart from interests, groups or personal information, Second Life users are able to specify regions on their profiles. The purpose of such favored regions of users is to let others know about their preferred locations. The following features are based on these regions and the types of measures are again the same as from the recorded regions. The favored regions of a user u were defined as $\Xi(u)$.

- *Common Favored Regions*: The number of favored regions two users u and v have in common was defined as $RF_C(u,v) = |\Xi(u) \cap \Xi(v)|$.
- *Total Favored Regions*: The total number of favored regions of two users u and v was formally stated as $RF_T(u,v) = |\Xi(u) \cup \Xi(v)|$.
- *Jaccard's Coefficient of Favored Regions*: The Jaccard's coefficient of the favored regions of two users u and v was written as $RF_{JC}(u,v) = \frac{|\Xi(u) \cap \Xi(v)|}{|\Xi(u) \cup \Xi(v)|}$.
- *Overlap of Favored Regions*: This feature represents the overlap between the common favored regions of two users u and v and the sum of the favored regions of u and the favored regions of v as $RF_O(u,v) = \frac{|\Xi(u) \cap \Xi(v)|}{|\Xi(u)| + |\Xi(v)|}$.

4.2 Location-Based Social Network Features

In this section the features derived from the location-based social network data are presented.

Topological Features

In the location-based social network the neighbors of a user $u \in V_L$ were defined as $\Gamma(u) = \{v \mid (u,v) \in E_L\}$. Similar to the topological online social network features described in Section 4.1, features to measure the structural

overlap of two users in the location-based social network were subdivided as follows:

- *Common Neighbors*: This feature represents the number of neighbors two users u and v have in common. The common neighbors were denoted as $L_{CN}(u, v) = |\Gamma(u) \cap \Gamma(v)|$.
- *Total Neighbors*: As already mentioned the number of total neighbors of two users u and v is the length of the union set of the neighbors of both of them and was defined as $L_{TN}(u, v) = |\Gamma(u) \cup \Gamma(v)|$.
- *Jaccard's Coefficient*: The Jaccard's coefficient for two users u and v in the location-based social network was stated as $L_{JC}(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$.
- *Adamic Adar*: Slightly different from the Adamic Adar measure of the online social network described in Section 4.1 the Adamic Adar for undirected networks was formally defined as $L_{AA}(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{log(|\Gamma(z)|)}$.

Homophilic Features

As mentioned in Section 3.3 the implemented robots monitored users in-world at Second Life events. The events a user u visited were defined as $\Phi(u)$. The following features refer to such events and their locations and were taken from Steurer and Trattner [33] and Cranshaw et al. [11]:

- *Common Events*: The number of common events two users u and v visited was defined as $E_C(u, v) = |\Pi(u) \cap \Pi(v)|$.
- *Total Events*: The number of all events joined by two users u and v was formally stated as $E_T(u, v) = |\Pi(u) \cup \Pi(v)|$.
- *Jaccard's Coefficient of Events*: The Jaccard's coefficient measure of the events two users u and v visited could be computed as $E_{JC}(u, v) = \frac{|\Pi(u) \cap \Pi(v)|}{|\Pi(u) \cup \Pi(v)|}$.
- *Cosine Similarity of Event Categories*: Another way to measure the similarity between two users u and v is to compute the cosine similarity of two vectors including some user specific attributes. In this case two vectors $\delta(u)$ and $\delta(v)$ with the length of the number of all categories of

the Second Life events for each user pair (u, v) were defined. Every item i in such a vector represented the number of events the user visited of a specific category. The cosine similarity of event categories between two users u and v could now be computed as $E_{CCos}(u, v) = \frac{\delta(u) \cdot \delta(v)}{\|\delta(u)\| \ \|\delta(v)\|}$.

The information of the following features is based on the regions of the visited events of the users. The measures of the features are the same as from the recorded and favored regions:

- *Common Event Regions*: The number of regions of events two users u and v visited in common were stated as $RE_C(u, v) = |\Upsilon(u) \cap \Upsilon(v)|$.
- *Total Event Regions*: The total number of event regions of two users u and v were formally denoted as $RE_T(u, v) = |\Upsilon(u) \cup \Upsilon(v)|$.
- *Jaccard's Coefficient of Event Regions*: This feature measures the Jaccard's coefficient of the event regions of two users u and v: $RE_{JC}(u, v) = \frac{|\Upsilon(u) \cap \Upsilon(v)|}{|\Upsilon(u) \cup \Upsilon(v)|}$.
- *Overlap of Event Regions*: The overlap between the common event regions of two users u and v and the sum of the event regions of u and the event regions of v were defined in this feature as $RE_O(u, v) = \frac{|\Upsilon(u) \cap \Upsilon(v)|}{|\Upsilon(u)| + |\Upsilon(v)|}$.

4.3 Trading Network Features

This section provides a description of the features derived from the crawled trading network data.

Topological Features

The topological features to measure the structural overlap of two users in the online social network described in Section 4.1 could also be applied on the trading network of the Second Life Marketplace. Since this network is directed, some of the features could be split into outgoing and incoming features again. The outgoing neighbors in the trading network of a user $u \in V_T$ were denoted as $\Psi^+(u) = \{v \mid (u, v) \in E_T\}$ and incoming neighbors as $\Psi^-(u) = \{v \mid (v, u) \in E_T\}$. The formally definition of the whole set of neighbors was stated as $\Psi(u) = \Psi^+(u) \cup \Psi^-(u)$.

- *Common Outgoing Neighbors*: The number of outgoing neighbors two users u and v have in common was defined as $T^+_{CN}(u,v) = |\Psi^+(u) \cap \Psi^+(v)|$.

- *Common Incoming Neighbors*: The definition for the number of common incoming neighbors of two users u and v was stated as $T^-_{CN}(u,v) = |\Psi^-(u) \cap \Psi^-(v)|$.

- *Total Outgoing Neighbors*: The number of total outgoing neighbors of two users u and v in the trading network could be computed as $T^+_{TN}(u,v) = |\Psi^+(u) \cup \Psi^+(v)|$.

- *Total Incoming Neighbors*: The opposite to the total outgoing neighbors are the total incoming neighbors of two users u and v and was defined as $T^-_{TN}(u,v) = |\Psi^-(u) \cup \Psi^-(v)|$.

- *Outgoing Jaccard's Coefficient*: The definition of the outgoing Jaccard's coefficient of two users u and v of the trading network was denoted as $T^+_{JC}(u,v) = \frac{|\Psi^+(u) \cap \Psi^+(v)|}{|\Psi^+(u) \cup \Psi^+(v)|}$.

- *Incoming Jaccard's Coefficient*: The incoming Jaccard's coefficient is the complement to the previous feature and was written as $T^-_{JC}(u,v) = \frac{|\Psi^-(u) \cap \Psi^-(v)|}{|\Psi^-(u) \cup \Psi^-(v)|}$.

- *Preferential Attachment Out*: As mentioned above the preferential attachment score is a measure for the correlation between the out-degree of a user u and the in-degree of a user v and was stated as $T^+_{PS}(u,v) = |\Psi^+(u)| \cdot |\Psi^-(v)|$.

- *Preferential Attachment In*: The difference to the preferential attachment out feature is the swapping of the users. This feature was formally defined as $T^-_{PS}(u,v) = |\Psi^-(u)| \cdot |\Psi^+(v)|$.

- *Reciprocity of User Communication*: As already mentioned, the value for the reciprocity between two users u and v in a directed network is 1 if there is an edge in both directions and 0 if there is no bidirectional link between these users. Formally, this feature could be written as $T_R(u,v) = \begin{cases} 0 & \text{if } (u,v) \in E_T,\ (v,u) \notin E_T \\ 1 & \text{if } (u,v) \in E_T,\ (v,u) \in E_T \end{cases}$.

- *Adamic Adar*: Similar to the Adamic Adar measure for the online social network, this metric could also be used for the directed trading network as $T_{AA}(u,v) = \sum_{z \in \Psi^-(u) \cap \Psi^-(v)} \frac{1}{log(|\Psi(z)^-|)}$.

Homophilic Features

All homophilic features of the trading network of the Second Life Marketplace are based on the attributes of the traded products. The attributes are category, price and ratings of the products. The cosine similarity measures for the following features were taken again from Steurer and Trattner [33]:

- *Cosine Similarity of Product Categories*: To compute a value for the similarity between the product categories of a user pair (u,v) two vectors $\zeta(u)$ and $\zeta(v)$ were defined. The vectors' lengths were the number of all product categories of the products u and v bought or sold. So each item i in these vectors represented a product category. The values for i were the number of products in a specific category that the user traded with. Similarly to the cosine similarity of event categories feature in the homophilic feature set of the location-based social network in Section 4.2 the cosine similarity of product categories between u and v could be computed as $P_{CCos}(u,v) = \frac{\zeta(u) \cdot \zeta(v)}{\|\zeta(u)\| \ \|\zeta(v)\|}$.

- *Cosine Similarity of Product Prices*: The same metric could be applied for product prices. The prices were graduated by the following scheme: $0-5L\$$, $6-10L\$$, $11-20L\$$, $21-50L\$$, $51-200L\$$, $201-500L\$$, $501L\$-\infty$. The vectors with the number of products per price step for two users u and v were denoted as $\rho(u)$ and $\rho(v)$ and so the cosine similarity of product prices between u and v could formally be written as $P_{PCos}(u,v) = \frac{\rho(u) \cdot \rho(v)}{\|\rho(u)\| \ \|\rho(v)\|}$.

- *Cosine Similarity of Product Ratings*: The cosine similarity could also be calculated for the user ratings of the products. The products were classified in ten different rating schemes from 0.0 to 5.0 in incremental steps of 0.5. Each item i of the two vectors $\tau(u)$ and $\tau(v)$ of the users u and v represented the number of traded products by u and v in each product rating class. So the value of this feature could be computed as $P_{RCos}(u,v) = \frac{\tau(u) \cdot \tau(v)}{\|\tau(u)\| \ \|\tau(v)\|}$.

4.4 Overview of all Features

Table 4.1 in this section gives a clear overview of the overall 50 used features consisting of online social network features, location-based social network features and trading network features, each set split into topological and homophilic features.

Table 4.1: Overview of all features.

		Feature	Description	Formal Definition
Online Social Network	Topological	O^+_{CN}	Common Outgoing Neighbors	$O^+_{CN}(u,v) = \lvert\Theta^+(u) \cap \Theta^+(v)\rvert$
		O^-_{CN}	Common Incoming Neighbors	$O^-_{CN}(u,v) = \lvert\Theta^-(u) \cap \Theta^-(v)\rvert$
		O^+_{TN}	Total Outgoing Neighbors	$O^+_{TN}(u,v) = \lvert\Theta^+(u) \cup \Theta^+(v)\rvert$
		O^-_{TN}	Total Incoming Neighbors	$O^-_{TN}(u,v) = \lvert\Theta^-(u) \cup \Theta^-(v)\rvert$
		O^+_{JC}	Outgoing Jaccard's Coefficient	$O^+_{JC}(u,v) = \frac{\lvert\Theta^+(u) \cap \Theta^+(v)\rvert}{\lvert\Theta^+(u) \cup \Theta^+(v)\rvert}$
		O^-_{JC}	Incoming Jaccard's Coefficient	$O^-_{JC}(u,v) = \frac{\lvert\Theta^-(u) \cap \Theta^-(v)\rvert}{\lvert\Theta^-(u) \cup \Theta^-(v)\rvert}$
		O^+_{PS}	Preferential Attachment Out	$O^+_{PS}(u,v) = \lvert\Theta^+(u)\rvert \cdot \lvert\Theta^-(v)\rvert$
		O^-_{PS}	Preferential Attachment In	$O^-_{PS}(u,v) = \lvert\Theta^-(u)\rvert \cdot \lvert\Theta^+(v)\rvert$
		O_R	Reciprocity of User Communication	$O_R(u,v) = \begin{cases} 0 & \text{if } (u,v) \in E_O,\ (v,u) \notin E_O \\ 1 & \text{if } (u,v) \in E_O,\ (v,u) \in E_O \end{cases}$
		O_{AA}	Adamic Adar	$O_{AA}(u,v) = \sum_{z\ \in\ \Theta^-(u)\ \cap\ \Theta^-(v)} \frac{1}{log(\lvert\Theta(z)^-\rvert)}$
	Homophilic	G_C	Common Groups	$G_C(u,v) = \lvert\Delta(u) \cap \Delta(v)\rvert$
		G_T	Total Groups	$G_T(u,v) = \lvert\Delta(u) \cup \Delta(v)\rvert$
		G_{JC}	Jaccard's Coefficient of Groups	$G_{JC}(u,v) = \frac{\lvert\Delta(u) \cap \Delta(v)\rvert}{\lvert\Delta(u) \cup \Delta(v)\rvert}$
		I_C	Common Interests	$I_C(u,v) = \lvert\Phi(u) \cap \Phi(v)\rvert$
		I_T	Total Interests	$I_T(u,v) = \lvert\Phi(u) \cup \Phi(v)\rvert$
		I_{JC}	Jaccard's Coefficient of Interests	$I_{JC}(u,v) = \frac{\lvert\Phi(u) \cap \Phi(v)\rvert}{\lvert\Phi(u) \cup \Phi(v)\rvert}$
		OI	Interactions	$OI(u,v) = \lvert\iota(u,v)\rvert$
		RR_C	Common Recorded Regions	$RR_C(u,v) = \lvert\Lambda(u) \cap \Lambda(v)\rvert$
		RR_T	Total Recorded Regions	$RR_T(u,v) = \lvert\Lambda(u) \cup \Lambda(v)\rvert$
		RR_{JC}	Jaccard's Coefficient of Recorded Regions	$RR_{JC}(u,v) = \frac{\lvert\Lambda(u) \cap \Lambda(v)\rvert}{\lvert\Lambda(u) \cup \Lambda(v)\rvert}$
		RR_O	Overlap of Recorded Regions	$RR_O(u,v) = \frac{\lvert\Lambda(u) \cap \Lambda(v)\rvert}{\lvert\Lambda(u)\rvert + \lvert\Lambda(v)\rvert}$
		RF_C	Common Favored Regions	$RF_C(u,v) = \lvert\Xi(u) \cap \Xi(v)\rvert$
		RF_T	Total Favored Regions	$RF_T(u,v) = \lvert\Xi(u) \cup \Xi(v)\rvert$
		RF_{JC}	Jaccard's Coefficient of Favored Regions	$RF_{JC}(u,v) = \frac{\lvert\Xi(u) \cap \Xi(v)\rvert}{\lvert\Xi(u) \cup \Xi(v)\rvert}$
		RF_O	Overlap of Favored Regions	$RF_O(u,v) = \frac{\lvert\Xi(u) \cap \Xi(v)\rvert}{\lvert\Xi(u)\rvert + \lvert\Xi(v)\rvert}$
Location-Based Social Network	Top.	L_{CN}	Common Neighbors	$L_{CN}(u,v) = \lvert\Gamma(u) \cap \Gamma(v)\rvert$
		L_{TN}	Total Neighbors	$L_{TN}(u,v) = \lvert\Gamma(u) \cup \Gamma(v)\rvert$
		L_{JC}	Jaccard's Coefficient	$L_{JC}(u,v) = \frac{\lvert\Gamma(u) \cap \Gamma(v)\rvert}{\lvert\Gamma(u) \cup \Gamma(v)\rvert}$
		L_{AA}	Adamic Adar	$L_{AA}(u,v) = \sum_{z\ \in\ \Gamma(u)\ \cap\ \Gamma(v)} \frac{1}{log(\lvert\Gamma(z)\rvert)}$
	Homophilic	E_C	Common Events	$E_C(u,v) = \lvert\Pi(u) \cap \Pi(v)\rvert$
		E_T	Total Events	$E_T(u,v) = \lvert\Pi(u) \cup \Pi(v)\rvert$
		E_{JC}	Jaccard's Coefficient of Events	$E_{JC}(u,v) = \frac{\lvert\Pi(u) \cap \Pi(v)\rvert}{\lvert\Pi(u) \cup \Pi(v)\rvert}$
		E_{CCos}	Cosine Similarity of Event Categories	$E_{CCos}(u,v) = \frac{\delta(u) \cdot \delta(v)}{\lVert\delta(u)\rVert\ \lVert\delta(v)\rVert}$
		RE_C	Common Event Regions	$RE_C(u,v) = \lvert\Upsilon(u) \cap \Upsilon(v)\rvert$
		RE_T	Total Event Regions	$RE_T(u,v) = \lvert\Upsilon(u) \cup \Upsilon(v)\rvert$
		RE_{JC}	Jaccard's Coefficient of Event Regions	$RE_{JC}(u,v) = \frac{\lvert\Upsilon(u) \cap \Upsilon(v)\rvert}{\lvert\Upsilon(u) \cup \Upsilon(v)\rvert}$
		RE_O	Overlap of Event Regions	$RE_O(u,v) = \frac{\lvert\Upsilon(u) \cap \Upsilon(v)\rvert}{\lvert\Upsilon(u)\rvert + \lvert\Upsilon(v)\rvert}$
Trading Network	Topological	T^+_{CN}	Common Outgoing Neighbors	$T^+_{CN}(u,v) = \lvert\Psi^+(u) \cap \Psi^+(v)\rvert$
		T^-_{CN}	Common Incoming Neighbors	$T^-_{CN}(u,v) = \lvert\Psi^-(u) \cap \Psi^-(v)\rvert$
		T^+_{TN}	Total Outgoing Neighbors	$T^+_{TN}(u,v) = \lvert\Psi^+(u) \cup \Psi^+(v)\rvert$
		T^-_{TN}	Total Incoming Neighbors	$T^-_{TN}(u,v) = \lvert\Psi^-(u) \cup \Psi^-(v)\rvert$
		T^+_{JC}	Outgoing Jaccard's Coefficient	$T^+_{JC}(u,v) = \frac{\lvert\Psi^+(u) \cap \Psi^+(v)\rvert}{\lvert\Psi^+(u) \cup \Psi^+(v)\rvert}$
		T^-_{JC}	Incoming Jaccard's Coefficient	$T^-_{JC}(u,v) = \frac{\lvert\Psi^-(u) \cap \Psi^-(v)\rvert}{\lvert\Psi^-(u) \cup \Psi^-(v)\rvert}$
		T^+_{PS}	Preferential Attachment Out	$T^+_{PS}(u,v) = \lvert\Psi^+(u)\rvert \cdot \lvert\Psi^-(v)\rvert$
		T^-_{PS}	Preferential Attachment In	$T^-_{PS}(u,v) = \lvert\Psi^-(u)\rvert \cdot \lvert\Psi^+(v)\rvert$
		T_R	Reciprocity of User Communication	$T_R(u,v) = \begin{cases} 0 & \text{if } (u,v) \in E_T,\ (v,u) \notin E_T \\ 1 & \text{if } (u,v) \in E_T,\ (v,u) \in E_T \end{cases}$
		T_{AA}	Adamic Adar	$T_{AA}(u,v) = \sum_{z\ \in\ \Psi^-(u)\ \cap\ \Psi^-(v)} \frac{1}{log(\lvert\Psi(z)^-\rvert)}$
	Hom.	P_{CCos}	Cosine Similarity of Product Categories	$P_{CCos}(u,v) = \frac{\zeta(u) \cdot \zeta(v)}{\lVert\zeta(u)\rVert\ \lVert\zeta(v)\rVert}$
		P_{PCos}	Cosine Similarity of Product Prices	$P_{PCos}(u,v) = \frac{\rho(u) \cdot \rho(v)}{\lVert\rho(u)\rVert\ \lVert\rho(v)\rVert}$
		P_{RCos}	Cosine Similarity of Product Ratings	$P_{RCos}(u,v) = \frac{\tau(u) \cdot \tau(v)}{\lVert\tau(u)\rVert\ \lVert\tau(v)\rVert}$

5 Evaluation Methodology

In the previous chapters the different sources of data and the various features for each network were described. In this chapter the setup for the experiments used in this thesis to answer the research questions (see Chapter 1) is described. To answer the first four research questions, the features of each of the networks were carefully separately examined. With this measure it was possible to give a comparison of the three networks' features and feature sets related to the prediction of trading interactions between user pairs. Research question five is about how the features of the online social network and the location-based social network perform to predict trading relations in comparison to the features of the trading network. And at long last the question about to what extent the combinations of the features of the three networks perform has to be posed.

To make the results for the different networks comparable, it was necessary to bring them on a common basis. So the online social network, the location-based social network and the trading network were intersected by picking out the common nodes of all networks. This means that only those users were considered who were active in all of the different networks, thus there was information about them in all of the network sources. Therefore, each user must have made at least one purchase as seller or buyer in the Second Life Marketplace, one interaction on My Second Life and one in-world observation by the robots.

This measure resulted in a remaining amount of data of $3,141$ users with $1,271$ interaction edges from the online social network, $8,462$ observing edges from the location-based social network and $2,748$ purchase interactions

Table 5.1: Statistics of the combined network.

Combined Network G_C	
Users	3,141
Online Social Network Edges	1,271
Location-Based Social Network Edges	8,462
Trading Network Edges	2,748
Total Edges	12,250
Degree	9.05
Connected Components	17
Largest Connected Component	98.42%
Sellers	914
Buyers	2,776
Sellers + Buyers	549

edges from the trading network. The total number of edges in the combined network was 12,250. There were 914 sellers and 2,776 buyers left, whereof 549 users were both sellers and buyers. The basic network statistics for this combined network are displayed in Table 5.1.

This combined network was formally defined as $G_C = \langle V_C, E_C \rangle$, where V_C was the set of common users of the three networks, the online social network G_O, the location-based social network G_L and the trading network G_T: $V_C = \{u \mid u \in V_O,\ u \in V_L,\ u \in V_T\}$. E_C was the union set of edges representing the relations between these users in either networks: $E_C = \{(u, v) \mid (u, v) \in E_O \text{ or } (u, v) \in E_L \text{ or } (u, v) \in E_T, \text{ and } u, v \in V_C\}$.

5.1 Implementation

All basic computations for preparing the experiments were done with Python. One reason for selecting this programming language was the extensive library *networkx*[1], which was very useful to handle the different types of networks.

First of all, the crawled data of the three network sources were extracted from the MySQL database and then the networks were established with methods of the networkx library. The nodes – representing the users – and the edges – showing a relation between users – in these networks had appropriate attributes, which were used for further feature calculations. The next step was combining the networks on the basis of common users to get

[1] `http://networkx.github.io/`

an initial point for the experiments.

All experiments were completed in two slightly different ways. The idea for the first way was that the starting point is a random seller s. The prediction result should tell to what extent any random buyer b will buy from s based on appropriate features. The second way was the other way around. The starting point is a random buyer b and the prediction result of the experiment should tell to what extent b will buy from any random seller s based on appropriate features. The difference from a technical point of view is that the direction of the trading network changes. First, the edges in the network go from sellers to buyers (Seller $\longrightarrow$ Buyer) and then the edges go from buyers to sellers (Seller $\longleftarrow$ Buyer).

5.2 Evaluation Approach

As highlighted in Section 2.2 basically two different approaches are commonly used to obtain a probability value for the prediction of trading relations between user pairs:

5.2.1 Supervised Learning

For the machine learning task called *Supervised Learning*, the suggestion of Guha et al. [15] to create a balanced dataset of user pairs with and without purchases were followed. $2,000$ user pairs of the combined network were randomly chosen, whereof $1,000$ had a purchase relation and $1,000$ had not. To bring this binary classification onto a common basis, all chosen user pairs had to consist of a seller and a buyer. With this rule it was prevented to select a user pair consisting of, for example, two buyers and make a purchase prediction for them, which would not have made sense. These $2,000$ user pairs were split into a training set to determine characteristics of purchase interactions and a test set for verification. This balanced sample of data resulted in a baseline of 50% for the trading prediction task when guessing at random. Finally, it was imported into the WEKA machine learning software for the prediction computation.

For all experiments in this thesis *logistic regression* with a tenfold cross-validation was used as main learning method. This is in line with the related work of Steurer and Trattner [33], who reached the best result with this method for their similar prediction problem (see Table 2.1).

5.2.2 Unsupervised Learning

The second approach to predict trading relations was *Unsupervised Learning* in terms of a *collaborative filtering* technique, which was first proposed by Liben et al. [25].

For each seller in the combined network the k-nearest neighbors for $k = 1, 3$ were computed. In a given data points collection a nearest neighbor of a query point is a data point that is closest to the query point [5]. This means for a network structure that the k-nearest neighbors for $k = 1$ of a node x is the set of directly connected neighbors of x. For example, for $k = 2$ all neighbors of x and further their neighbors would be considered. So a user ranking in terms of the value of a feature was calculated for each involved feature once for all buyers in the k-nearest neighbors set of a seller s and once for all buyers in the combined network. The *top ten buyers* of the ranked lists were now compared with a list of buyers who had a purchase relation with s. For each feature f for a seller s these comparisons resulted in a success rate as follows:

$$\textit{success-rate}(s, f) = \frac{|\textit{predicted-buyers}(s)|}{\textit{min}(|\textit{buyers}(s)|,\ 10)}.$$

The number of buyers of s found in the top ten list was divided by the minimum of the whole number of buyers of s and the list length ($\widehat{=}$ 10). Building the mean value of the success rates for all sellers in the network was a measure of how well a feature performed to predict trading relations. The overall success rate for all features was also computed by ranking the k-nearest neighbors including not just only one feature, but all involved features, comparing this list with the previously mentioned list of buyers for each seller and building the mean value. To make this result valid, all feature values were normalized, first. All users were considered for this process, regardless of whether user information needed for the current feature was available or not. For example, for the common groups feature of the online social network, users who had not declared any group in their profiles, were used.

5.3 Evaluation Metrics

To show the differences between user pairs with and without trading interactions the mean values of the features were computed by simply calculating

the average of the feature values of the involved edges. For the computation of the standard errors of the features, the *sem*[2] function of the *SciPy*[3] computing environment was used. The significance of each feature was calculated in several steps. First, the *Levene test* – introduced by Levene [24] – was done using the *levene*[4] function with all 1,000 positive and all 1,000 negative edges to test for equal variances. If the return variable *p_value* of this function was below 0.01, the *Wilcoxon rank-sum test* using the *ranksums*[5] function was done, otherwise the *two-sided Kolmogorov-Smirnof test* using the *ks_2samp*[6] function was executed. The returning *p_value* was the crucial measure for the significance of a feature. If this value was below 0.001, the feature was provided with three stars (***) in the appropriate tables. Two stars (**) for values below 0.01 and one star (*) for values below 0.1.

The correlation matrices illustrate the correlations between the involved features. They were computed using the *spearmanr*[7] function of the Python package *Statsmodels*[8] to calculate the Spearman rank-order correlation coefficient and the *p_value* for testing the significance. The significance levels were similar denoted as the significance levels for the mean values and standard errors: if the *p_value* was below 0.001, the feature pair was provided with three stars (***) in the appropriate tables. Two stars (**) for values below 0.01 and one star (*) for values below 0.1.

The information gain of each feature was computed with WEKA. It is a measure of how well a feature performs in terms of a trading interaction prediction. The values of the best features in each table were highlighted by way of illustration. For different features and feature sets the value of the area under the ROC curve (AUC) was generated using the WEKA machine learning software. Logistic regression as supervised learning method was applied. This value is a measure for the probability of the success rate for a

[2] http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.sem.html#scipy.stats.sem/

[3] http://scipy.org/

[4] http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.levene.html#scipy.stats.levene/

[5] http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ranksums.html#scipy.stats.ranksums/

[6] http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html#scipy.stats.ks_2samp/

[7] http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html#scipy.stats.spearmanr/

[8] http://pypi.python.org/pypi/statsmodels/

prediction of a trading interaction.

The tables in Section 6, which show the results of the unsupervised learning method collaborative filtering – described in Section 5.2 –, are always split in four value columns for the different k-values of the k-nearest neighbors.

6 Results

This chapter presents the results of the implemented experiments. Apart from the mean values and standard errors of the comparison between user pairs with and without trading interactions, the correlation matrices of the involved features, the values of the area under the ROC curve (AUC) for the different machine learning methods, the information gains and the computed success rates of the collaborative filtering are presented for each experiment.

6.1 Predicting Trading Interactions with Trading Network Features

In Table 6.1 the mean values and standard errors of the topological and homophilic features of the trading network are shown for user pairs with and without trading interactions. The significance of the features is also illustrated with stars. The feature $T^{+}_{PS}(u, v)$, which represents the outgoing preferential attachment score between two users u and v, shows considerable differences. The mean value for user pairs with one or more trading interactions was approximately 13 times higher as for users without a trading interaction. The number of common neighbors of two users $T^{+}_{CN}(u, v)$ and $T^{-}_{CN}(u, v)$ was about seven times higher for user pairs with trading interactions. A slightly less distinct was observed for the cosine similarity of the product categories of two users $P_{CCos}(u, v)$ with a barely higher value for user pairs with trading interactions.

Comparing topological features with homophilic features, the correlation matrices in Table 6.2 illustrate high correlations between the cosine similarity

Table 6.1: The mean values, standard errors and the significance of the trading network features to show the differences between user pairs with and without trading interactions.

			Seller ⟶ Buyer			Seller ⟵ Buyer		
			Trading Interactions			Trading Interactions		
Features			Yes	No	Sign.	Yes	No	Sign.
Trading Network	Topological	$T^{+}_{CN}(u,v)$	0.06 ± 0.03	0.01 ± 0.00		0.11 ± 0.01	0.01 ± 0.00	*
		$T^{-}_{CN}(u,v)$	0.12 ± 0.02	0.02 ± 0.00	*	0.05 ± 0.02	0.01 ± 0.01	
		$T^{+}_{TN}(u,v)$	51.55 ± 1.75	15.76 ± 0.89	***	25.70 ± 1.06	11.32 ± 0.47	***
		$T^{-}_{TN}(u,v)$	25.92 ± 1.15	11.17 ± 0.41	***	53.72 ± 1.86	15.24 ± 0.90	***
		$T^{+}_{PS}(u,v)$	1062.05 ± 87.31	84.50 ± 7.43	***	1069.64 ± 80.82	77.32 ± 6.53	***
		$T^{-}_{PS}(u,v)$	12.63 ± 2.95	13.05 ± 3.41		9.68 ± 2.17	17.33 ± 4.93	
		$T_{R}(u,v)$	0.01 ± 0.00	0.00 ± 0.00		0.01 ± 0.00	0.00 ± 0.00	
		$T_{AA}(u,v)$	0.07 ± 0.01	0.01 ± 0.01		0.01 ± 0.00	0.00 ± 0.00	
	Hom.	$P_{CCos}(u,v)$	0.52 ± 0.01	0.27 ± 0.01	***	0.49 ± 0.01	0.27 ± 0.01	***
		$P_{PCos}(u,v)$	0.55 ± 0.01	0.43 ± 0.01	***	0.55 ± 0.01	0.42 ± 0.01	***
		$P_{RCos}(u,v)$	0.81 ± 0.01	0.70 ± 0.01	***	0.81 ± 0.01	0.71 ± 0.01	***

of the product categories or ratings $P_{CCos}(u,v)$, $P_{RCos}(u,v)$ and the number of total neighbors or the outgoing preferential attachment score $T^{+}_{TN}(u,v)$, $T^{-}_{TN}(u,v)$, $T^{+}_{PS}(u,v)$ of two users u and v.

As Table 6.3 shows, for the prediction of a trading interaction of two users u and v, the best performing trading network features were the number of total neighbors $T^{+}_{TN}(u,v)$, $T^{-}_{TN}(u,v)$, the outgoing preferential attachment score $T^{+}_{PS}(u,v)$ and the cosine similarity of the product categories $P_{CCos}(u,v)$. In the first case (Seller ⟶ Buyer) the number of outgoing total neighbors feature $T^{+}_{TN}(u,v)$ was performing with an AUC value nearly 79% and very high success rates with the collaborative filtering method, whereas in the second case (Seller ⟵ Buyer) the number of incoming total neighbors feature $T^{-}_{TN}(u,v)$ had almost the same results. A reason for this observation is that with the turn of the direction in the network, the incoming and outgoing feature measures also change their places.

The information gain results of the features were quite similar to the AUC results and the success rates of the collaborative filtering. The best performance by far for a trading interaction prediction between two users u and v had the outgoing preferential attachment score $T^{+}_{PS}(u,v)$. Also acceptable information gains as well as success rates were observed for the total neighbors $T^{+}_{TN}(u,v)$ and $T^{-}_{TN}(u,v)$ and the cosine similarity of the product categories $P_{CCos}(u,v)$.

The topological feature set of the trading network was performing much

Table 6.2: Spearman's correlation matrices of the trading network features with their significance. The best correlations (> 0.20) between feature pairs of topological and homophilic features were highlighted.

(a) Seller $\longrightarrow$ Buyer

Features			Trading Network												
			Topological										Homophilic		
			T^+_{CN}	T^-_{CN}	T^+_{TN}	T^-_{TN}	T^+_{JC}	T^-_{JC}	T^+_{PS}	T^-_{PS}	T_R	T_{AA}	P_{CCos}	P_{PCos}	P_{RCos}
Trading Network	Topological	T^+_{CN}	1.00												
		T^-_{CN}	0.16^{***}	1.00											
		T^+_{TN}	0.16^{***}	0.09^{***}	1.00										
		T^-_{TN}	0.07^{**}	0.21^{***}	0.22^{***}	1.00									
		T^+_{JC}	1.00^{***}	0.16^{***}	0.16^{***}	0.07^{**}	1.00								
		T^-_{JC}	0.16^{***}	1.00^{***}	0.09^{***}	0.21^{***}	0.16^{***}	1.00							
		T^+_{PS}	0.13^{***}	0.16^{***}	0.82^{***}	0.56^{***}	0.13^{***}	0.16^{***}	1.00						
		T^-_{PS}	0.30^{***}	0.12^{***}	0.21^{***}	0.20^{***}	0.30^{***}	0.12^{***}	0.09^{***}	1.00					
		T_R	0.21^{***}	0.08^{***}	0.01	0.01	0.21^{***}	0.08^{***}	-0.03	0.16^{***}	1.00				
		T_{AA}	0.07^{**}	0.61^{***}	0.04^{*}	0.16^{***}	0.07^{**}	0.61^{***}	0.09^{***}	0.04^{*}	0.06^{**}	1.00			
	Hom.	P_{CCos}	0.07^{**}	0.09^{***}	$\mathbf{0.26^{***}}$	$\mathbf{0.26^{***}}$	0.06^{**}	0.09^{***}	$\mathbf{0.37^{***}}$	0.00	0.04^{*}	0.10^{***}	1.00		
		P_{PCos}	0.03	0.09^{***}	0.18^{***}	0.17^{***}	0.03	0.09^{***}	$\mathbf{0.27^{***}}$	0.04^{*}	0.01	0.05^{*}	0.20^{***}	1.00	
		P_{RCos}	0.04^{*}	0.10^{***}	$\mathbf{0.23^{***}}$	0.18^{***}	0.04^{*}	0.10^{***}	$\mathbf{0.27^{***}}$	0.05^{*}	0.07^{**}	0.07^{**}	0.17^{***}	0.14^{***}	1.00

(b) Seller $\longleftarrow$ Buyer

Features			Trading Network												
			Topological										Homophilic		
			T^+_{CN}	T^-_{CN}	T^+_{TN}	T^-_{TN}	T^+_{JC}	T^-_{JC}	T^+_{PS}	T^-_{PS}	T_R	T_{AA}	P_{CCos}	P_{PCos}	P_{RCos}
Trading Network	Topological	T^+_{CN}	1.00												
		T^-_{CN}	0.06^{*}	1.00											
		T^+_{TN}	0.24^{***}	0.11^{***}	1.00										
		T^-_{TN}	0.08^{***}	0.16^{***}	0.26^{***}	1.00									
		T^+_{JC}	1.00^{***}	0.05^{*}	0.24^{***}	0.08^{***}	1.00								
		T^-_{JC}	0.05^{*}	1.00^{***}	0.11^{***}	0.16^{***}	0.05^{*}	1.00							
		T^+_{PS}	0.16^{***}	0.14^{***}	0.57^{***}	0.84^{***}	0.16^{***}	0.14^{***}	1.00						
		T^-_{PS}	0.10^{***}	0.24^{***}	0.19^{***}	0.19^{***}	0.10^{***}	0.24^{***}	0.10^{***}	1.00					
		T_R	-0.02	0.23^{***}	0.03	0.03	-0.02	0.22^{***}	-0.01	0.21^{***}	1.00				
		T_{AA}	0.02	0.60^{***}	0.06^{**}	0.10^{***}	0.02	0.60^{***}	0.08^{***}	0.18^{***}	0.19^{***}	1.00			
	Hom.	P_{CCos}	0.13^{***}	0.10^{***}	$\mathbf{0.28^{***}}$	$\mathbf{0.27^{***}}$	0.13^{***}	0.10^{***}	$\mathbf{0.36^{***}}$	0.02	0.04^{*}	0.08^{***}	1.00		
		P_{PCos}	0.06^{**}	0.08^{***}	0.18^{***}	0.18^{***}	0.06^{**}	0.08^{***}	$\mathbf{0.26^{***}}$	0.03	-0.00	0.05^{*}	0.18^{***}	1.00	
		P_{RCos}	0.12^{***}	0.09^{***}	$\mathbf{0.26^{***}}$	$\mathbf{0.25^{***}}$	0.12^{***}	0.09^{***}	$\mathbf{0.31^{***}}$	0.07^{**}	0.04^{*}	0.04	0.15^{***}	0.19^{***}	1.00

better than the homophilic feature set. As the AUC values show, for a seller s, only considering the topological features, up to an extent of 83.60% can be predicted, if a random buyer b will buy from s. With nearly the same probability of 85.40% it can be predicted if a given buyer b will buy from a random seller s. The AUC values for the homophilic feature set of the trading network were in the range from about 72% to 73%.

It could be said that the homophilic features of the trading network reach a quite acceptable prediction result, but in combination with the topological features, they do not effect a considerable increase, because the overall result is quite the same as with the topological features alone: 84.00% or 84.90%.

Table 6.3: Supervised learning including the AUC values and the information gains and unsupervised learning with collaborative filtering for the trading network features. The best AUC values (> 70.00%), the best three information gains and the best success rates were highlighted.

(a) Seller ⟶ Buyer

			Supervised		Unsupervised		
		Features	AUC (LR)	InfoGain	SR@$k=1$	SR@$k=3$	SR@all
Trading Network	Topological	$T^{+}_{CN}(u,v)$	50.50%	< 0.01	0.1273	0.0497	0.0114
		$T^{-}_{CN}(u,v)$	51.80%	0.01316	0.2008	0.0696	0.0195
		$T^{+}_{TN}(u,v)$	**78.50%**	**0.19419**	**0.9110**	**0.2760**	0.0037
		$T^{-}_{TN}(u,v)$	69.70%	0.10363	**0.9521**	**0.3705**	**0.0521**
		$T^{+}_{JC}(u,v)$	50.50%	< 0.01	0.1273	0.0507	0.0130
		$T^{-}_{JC}(u,v)$	51.90%	0.01316	0.2003	0.0630	0.0084
		$T^{+}_{PS}(u,v)$	**84.80%**	**0.31210**	**0.9523**	**0.3706**	**0.0522**
		$T^{-}_{PS}(u,v)$	48.50%	< 0.01	0.4719	0.0546	0.0021
		$T_{R}(u,v)$	50.10%	< 0.01	0.2073	0.0439	0.0220
		$T_{AA}(u,v)$	50.70%	< 0.01	0.1346	0.0426	0.0114
	Topological		**83.60%**	—	**0.9415**	**0.3434**	**0.0338**
	Hom.	$P_{CCos}(u,v)$	**71.10%**	**0.11498**	**0.9067**	**0.2737**	0.0077
		$P_{PCos}(u,v)$	61.50%	0.04965	0.8911	0.2577	0.0003
		$P_{RCos}(u,v)$	63.10%	0.04319	0.8803	0.2556	0.0025
	Homophilic		**73.20%**	—	0.8207	0.1839	0.0000
Trading Network			**84.00%**	—	0.8422	0.2331	0.0244

(b) Seller ⟵ Buyer

			Supervised		Unsupervised		
		Features	AUC (LR)	InfoGain	SR@$k=1$	SR@$k=3$	SR@all
Trading Network	Topological	$T^{+}_{CN}(u,v)$	52.20%	0.02210	0.1447	0.1636	0.0278
		$T^{-}_{CN}(u,v)$	50.50%	< 0.01	0.0461	0.0324	0.0125
		$T^{+}_{TN}(u,v)$	**71.10%**	**0.10489**	**0.9929**	**0.3735**	0.0100
		$T^{-}_{TN}(u,v)$	**79.10%**	**0.21204**	**0.9974**	**0.6395**	**0.1287**
		$T^{+}_{JC}(u,v)$	52.20%	0.01920	0.1447	0.1661	0.0268
		$T^{-}_{JC}(u,v)$	50.60%	< 0.01	0.0461	0.0319	0.0087
		$T^{+}_{PS}(u,v)$	**86.20%**	**0.34578**	**0.9975**	**0.6395**	**0.1291**
		$T^{-}_{PS}(u,v)$	48.40%	< 0.01	0.1609	0.0323	0.0008
		$T_{R}(u,v)$	50.10%	< 0.01	0.0586	0.0185	0.0119
		$T_{AA}(u,v)$	50.10%	< 0.01	0.0175	0.0147	0.0061
	Topological		**85.40%**	—	**0.9967**	**0.4720**	**0.0683**
	Hom.	$P_{CCos}(u,v)$	69.40%	0.09929	**0.9528**	**0.2898**	0.0023
		$P_{PCos}(u,v)$	62.20%	0.04819	**0.9740**	**0.2824**	0.0034
		$P_{RCos}(u,v)$	62.70%	0.05176	**0.9749**	**0.3184**	0.0027
	Homophilic		**71.80%**	—	**0.9849**	0.2184	0.0002
Trading Network			**84.90%**	—	**0.9893**	**0.2991**	**0.0371**

Table 6.4: The mean values, standard errors and the significance of the online social network features to show the differences between user pairs with and without trading interactions.

			Seller ⟶ Buyer			Seller ⟵ Buyer		
			Trading Interactions			Trading Interactions		
		Features	Yes	No	Sign.	Yes	No	Sign.
Online Social Network	Topological	$O^+_{CN}(u,v)$	0.02 ± 0.01	0.04 ± 0.04		0.04 ± 0.03	0.00 ± 0.00	
		$O^-_{CN}(u,v)$	0.02 ± 0.01	0.00 ± 0.00		0.08 ± 0.08	0.00 ± 0.00	
		$O^+_{TN}(u,v)$	7.92 ± 0.72	7.39 ± 0.76		8.05 ± 0.76	8.27 ± 0.89	
		$O^-_{TN}(u,v)$	9.81 ± 0.86	8.17 ± 0.84	**	9.11 ± 0.90	8.91 ± 0.98	***
		$O^+_{PS}(u,v)$	11.39 ± 2.84	31.48 ± 18.20	***	38.02 ± 14.15	15.84 ± 2.60	***
		$O^-_{PS}(u,v)$	32.41 ± 8.12	13.65 ± 2.74	***	20.22 ± 10.05	8.46 ± 1.55	*
		$O_R(u,v)$	0.01 ± 0.00	0.00 ± 0.00		0.01 ± 0.00	0.00 ± 0.00	
		$O_{AA}(u,v)$	0.02 ± 0.01	0.00 ± 0.00		0.05 ± 0.05	0.00 ± 0.00	
	Homophilic	$G_C(u,v)$	0.17 ± 0.03	0.05 ± 0.01		0.14 ± 0.02	0.06 ± 0.01	
		$G_T(u,v)$	29.82 ± 0.59	32.84 ± 0.60	**	29.15 ± 0.60	32.18 ± 0.59	*
		$I_C(u,v)$	0.03 ± 0.01	0.01 ± 0.00		0.02 ± 0.01	0.01 ± 0.00	
		$I_T(u,v)$	6.65 ± 0.28	6.42 ± 0.26		6.65 ± 0.27	6.10 ± 0.26	
		$OI(u,v)$	0.07 ± 0.04	0.00 ± 0.00		1.37 ± 1.33	0.00 ± 0.00	
		$RR_C(u,v)$	0.00 ± 0.00	0.00 ± 0.00		0.01 ± 0.01	0.00 ± 0.00	
		$RR_T(u,v)$	2.52 ± 0.35	2.55 ± 0.24		2.56 ± 0.23	2.95 ± 0.28	
		$RF_C(u,v)$	0.02 ± 0.01	0.00 ± 0.00		0.03 ± 0.01	0.01 ± 0.00	
		$RF_T(u,v)$	5.99 ± 0.11	6.07 ± 0.11		5.90 ± 0.11	6.00 ± 0.11	

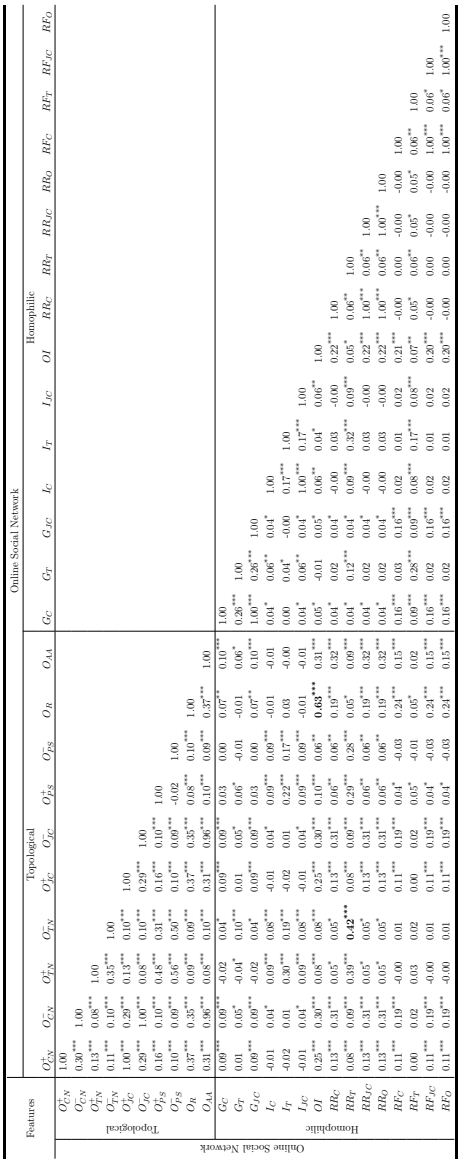
6.2 Predicting Trading Interactions with Online Social Network Features

The comparison of the feature values of the online social network between user pairs with and without trading interactions resulted in the conclusion, that the only mentionable significant differences were observed with the preferential attachment score features $O^+_{PS}(u,v)$, $O^-_{PS}(u,v)$, as shown in Table 6.4. The mean values lay about 2.5 times apart.

The correlation matrices for the features of the online social network in Table 6.5 show considerable differences for the two versions of the experiments with inverting the direction of the network. Comparing the topological with the homophilic features, for the first experiment (Seller ⟶ Buyer) high correlations between the number of total recorded regions $RR_T(u,v)$ and the number of total incoming neighbors $O^-_{TN}(u,v)$ of two users u and v and between the number of interactions $OI(u,v)$ and the reciprocity of the user communication $O_R(u,v)$ between u and v were observed. The second experiment (Seller ⟵ Buyer) resulted in much higher feature correlations be-

Table 6.5: Spearman's correlation matrices of the online social network features with their significance. The best correlations (> 0.40) between feature pairs of topological and homophilic features were highlighted.

(a) Seller $\longrightarrow$ Buyer



(b) Seller $\longleftarrow$ Buyer

tween topological and homophilic features of the online social network. The highest significant correlations were observed between the common neighbors, the Jaccard's coefficient, the reciprocity of the user communication or the Adamic Adar measure $O^{+}_{CN}(u,v)$, $O^{-}_{CN}(u,v)$, $O^{+}_{JC}(u,v)$, $O^{-}_{JC}(u,v)$,

$O_R(u,v)$, $O_{AA}(u,v)$ and the recorded region features $RR_C(u,v)$, $RR_{JC}(u,v)$, $RR_O(u,v)$ of two users u and v.

The computation of the AUC values, the information gains and the success rates with collaborative filtering of the online social network features yielded a moderate result, as shown in Table 6.6. For each individual feature, the best result was observed for the outgoing preferential attachment score $O^+_{PS}(u,v)$ between two users u and v with an AUC value of 56.60% and an information gain of 0.01169. Additionally, with unsupervised learning the total neighbors $T^+_{TN}(u,v)$, $T^-_{TN}(u,v)$, the total groups $G_T(u,v)$ and total favored regions $RF_T(u,v)$ performed well. Most of the other features reached nearly the same result as expected for flipping a coin: ≈50%.

However, with the combination of the features the results could be improved. The combination of all topological features reached an AUC value of 59.90% and the homophilic feature set was a little worse with 57.10%. Also plausible success rates could be recognized with the combination of topological and homophilic features. It could be observed that the topological features attained nearly the same result as all online social network features together. So the homophilic feature set in combination with the topological feature set did not increase the result over the baseline.

Table 6.6: Supervised learning including the AUC values and the information gains and unsupervised learning with collaborative filtering for the online social network features. The best AUC values (> 55.00%), the best two information gains and the best success rates were highlighted.

(a) Seller ⟶ Buyer

			Supervised		Unsupervised		
		Features	AUC (LR)	InfoGain	SR@$k=1$	SR@$k=3$	SR@all
Online Social Network	Topological	$O^{+}_{CN}(u,v)$	49.30%	< 0.01	0.0834	0.0326	0.0097
		$O^{-}_{CN}(u,v)$	49.90%	< 0.01	0.0963	0.0374	0.0138
		$O^{+}_{TN}(u,v)$	48.60%	< 0.01	**0.9084**	**0.3033**	0.0061
		$O^{-}_{TN}(u,v)$	54.30%	< 0.01	**0.9032**	**0.2906**	0.0068
		$O^{+}_{JC}(u,v)$	50.10%	< 0.01	0.0834	0.0329	0.0112
		$O^{-}_{JC}(u,v)$	49.40%	< 0.01	0.0963	0.0380	0.0164
		$O^{+}_{PS}(u,v)$	52.50%	< 0.01	0.4973	0.1663	0.0060
		$O^{-}_{PS}(u,v)$	**56.60%**	**0.01169**	0.7469	0.2375	0.0061
		$O_{R}(u,v)$	50.10%	< 0.01	0.1593	0.0510	0.0196
		$O_{AA}(u,v)$	50.00%	< 0.01	0.0733	0.0284	0.0107
		Topological	**59.90%**	—	**0.9058**	**0.3371**	0.0293
	Homophilic	$G_{C}(u,v)$	51.20%	< 0.01	0.4046	0.2114	**0.0505**
		$G_{T}(u,v)$	54.60%	< 0.01	**0.9069**	**0.2866**	0.0040
		$G_{JC}(u,v)$	51.30%	**0.01168**	0.4046	0.2115	**0.0586**
		$I_{C}(u,v)$	50.50%	< 0.01	0.1020	0.0327	0.0042
		$I_{T}(u,v)$	48.70%	< 0.01	0.8786	**0.2946**	0.0105
		$I_{JC}(u,v)$	50.50%	< 0.01	0.1020	0.0350	0.0054
		$OI(u,v)$	50.20%	< 0.01	0.0302	0.0159	0.0140
		$RR_{C}(u,v)$	49.70%	< 0.01	0.0277	0.0102	0.0022
		$RR_{T}(u,v)$	48.10%	< 0.01	0.8003	0.2502	0.0029
		$RR_{JC}(u,v)$	49.70%	< 0.01	0.0277	0.0102	0.0018
		$RR_{O}(u,v)$	49.70%	< 0.01	0.0277	0.0102	0.0018
		$RF_{C}(u,v)$	50.30%	< 0.01	0.1681	0.0889	**0.0414**
		$RF_{T}(u,v)$	48.90%	< 0.01	**0.9168**	**0.2959**	0.0086
		$RF_{JC}(u,v)$	50.30%	< 0.01	0.1681	0.0888	**0.0411**
		$RF_{O}(u,v)$	50.30%	< 0.01	0.1681	0.0888	**0.0411**
		Homophilic	**57.10%**	—	**0.9136**	**0.3228**	0.0197
Online Social Network			**59.80%**	—	**0.9084**	**0.3225**	0.0220

(b) Seller ⟵ Buyer

			Supervised		Unsupervised		
		Features	AUC (LR)	InfoGain	SR@$k=1$	SR@$k=3$	SR@all
Online Social Network	Topological	$O^{+}_{CN}(u,v)$	50.30%	< 0.01	0.0545	0.0203	0.0089
		$O^{-}_{CN}(u,v)$	49.90%	< 0.01	0.0596	0.0263	0.0111
		$O^{+}_{TN}(u,v)$	48.90%	< 0.01	**0.9362**	**0.3634**	0.0085
		$O^{-}_{TN}(u,v)$	50.70%	< 0.01	**0.9868**	**0.4215**	0.0116
		$O^{+}_{JC}(u,v)$	50.30%	< 0.01	0.0545	0.0204	0.0089
		$O^{-}_{JC}(u,v)$	49.90%	< 0.01	0.0598	0.0264	0.0124
		$O^{+}_{PS}(u,v)$	**56.60%**	**0.01649**	0.7063	**0.2999**	0.0090
		$O^{-}_{PS}(u,v)$	48.50%	< 0.01	0.6108	0.2569	0.0074
		$O_{R}(u,v)$	50.10%	< 0.01	0.0527	0.0181	0.0112
		$O_{AA}(u,v)$	49.90%	< 0.01	0.0490	0.0216	0.0104
		Topological	**58.00%**	—	**0.9917**	**0.4144**	0.0276
	Homophilic	$G_{C}(u,v)$	50.90%	< 0.01	0.2589	0.1790	**0.0485**
		$G_{T}(u,v)$	54.50%	< 0.01	**0.9830**	**0.3965**	0.0149
		$G_{JC}(u,v)$	51.00%	**0.01215**	0.2589	0.1797	**0.0537**
		$I_{C}(u,v)$	50.00%	< 0.01	0.0524	0.0488	0.0046
		$I_{T}(u,v)$	51.70%	< 0.01	0.7892	**0.3511**	0.0077
		$I_{JC}(u,v)$	50.00%	< 0.01	0.0524	0.0489	0.0051
		$OI(u,v)$	50.10%	< 0.01	0.0197	0.0152	0.0145
		$RR_{C}(u,v)$	49.70%	< 0.01	0.0219	0.0080	0.0033
		$RR_{T}(u,v)$	49.90%	< 0.01	0.6435	**0.3083**	0.0100
		$RR_{JC}(u,v)$	49.80%	< 0.01	0.0219	0.0080	0.0027
		$RR_{O}(u,v)$	49.80%	< 0.01	0.0219	0.0080	0.0027
		$RF_{C}(u,v)$	50.40%	< 0.01	0.1009	0.0618	**0.0309**
		$RF_{T}(u,v)$	50.50%	< 0.01	**0.9899**	**0.3827**	0.0059
		$RF_{JC}(u,v)$	50.40%	< 0.01	0.1009	0.0618	**0.0324**
		$RF_{O}(u,v)$	50.40%	< 0.01	0.1009	0.0618	**0.0324**
		Homophilic	**58.00%**	—	**0.9936**	**0.4160**	**0.0302**
Online Social Network			**59.30%**	—	**0.9926**	**0.4198**	**0.0316**

Table 6.7: The mean values, standard errors and the significance of the location-based social network features to show the differences between user pairs with and without trading interactions.

			Seller ⟶ Buyer			Seller ⟵ Buyer		
			Trading Interactions			Trading Interactions		
Features			Yes	No	Sign.	Yes	No	Sign.
Location-Based Social Network	Topological	$L_{CN}(u,v)$	1.32 ± 0.58	0.17 ± 0.08		1.61 ± 0.59	0.20 ± 0.05	
		$L_{TN}(u,v)$	169.42 ± 14.98	128.36 ± 7.98		157.24 ± 12.90	169.51 ± 16.07	*
		$L_{JC}(u,v)$	0.01 ± 0.00	0.00 ± 0.00		0.01 ± 0.00	0.00 ± 0.00	
		$L_{AA}(u,v)$	0.82 ± 0.40	0.07 ± 0.03		1.03 ± 0.41	0.09 ± 0.02	
	Homophilic	$E_C(u,v)$	0.30 ± 0.21	0.00 ± 0.00		0.15 ± 0.06	0.00 ± 0.00	
		$E_T(u,v)$	42.69 ± 3.56	31.41 ± 1.69		39.74 ± 3.03	42.70 ± 3.74	
		$E_{CCos}(u,v)$	0.51 ± 0.01	0.48 ± 0.01		0.48 ± 0.01	0.49 ± 0.01	
		$RE_C(u,v)$	0.21 ± 0.02	0.14 ± 0.01		0.21 ± 0.02	0.13 ± 0.01	*
		$RE_T(u,v)$	12.22 ± 0.37	12.02 ± 0.33		12.26 ± 0.37	12.46 ± 0.32	
		$RE_{JC}(u,v)$	0.02 ± 0.00	0.01 ± 0.00	*	0.02 ± 0.00	0.01 ± 0.00	*
		$RE_O(u,v)$	0.02 ± 0.00	0.01 ± 0.00	*	0.02 ± 0.00	0.01 ± 0.00	*

6.3 Predicting Trading Interactions with Location-Based Social Network Features

The differences of the feature mean values and standard errors of the location-based social network between user pairs with and without trading interactions are shown in Table 6.7. The highest observed significance for the location-based social network features was at level 0.1 (*). The common neighbors $L_{CN}(u,v)$ of two users u and v in the network with mean values of about 8 times higher for user pairs with trading interactions than without trading interactions and the Adamic Adar measure $L_{AA}(u,v)$ with mean values about 11.5 times higher were recognized as the features with the highest differences.

The highest correlation by far between the topological and the homophilic features of the location-based social network was observed between the number of total neighbors $L_{TN}(u,v)$ and the number of total events $E_T(u,v)$ of two users u and v with about 0.85^{***}. Also a high correlation with a significance at level 0.001 was recognized between the number of total neighbors $L_{TN}(u,v)$ and the number of total event regions $RE_T(u,v)$ with $\approx 0.58^{***}$.

As Table 6.9 shows, the performance of the features of the location-based social network for predicting trading interactions with supervised learning left much to be desired. Each feature on its own had coin flipping characteristics – the AUC values were around 50%. Only the information gain for

Table 6.8: Spearman's correlation matrices of the location-based social network features with their significance. The best correlations (> 0.40) between feature pairs of topological and homophilic features were highlighted.

(a) Seller $\longrightarrow$ Buyer

Features			Location-Based Social Network											
			Topological				Homophilic							
			L_{CN}	L_{TN}	L_{JC}	L_{AA}	E_C	E_T	E_{JC}	E_{CCos}	RE_C	RE_T	RE_{JC}	RE_O
Location-Based Social Network	Topological	L_{CN}	1.00											
		L_{TN}	0.23^{***}	1.00										
		L_{JC}	1.00^{***}	0.23^{***}	1.00									
		L_{AA}	1.00^{***}	0.23^{***}	1.00^{***}	1.00								
	Homophilic	E_C	$\mathbf{0.47^{***}}$	0.07^{**}	$\mathbf{0.47^{***}}$	$\mathbf{0.47^{***}}$	1.00							
		E_T	0.23^{***}	$\mathbf{0.86^{***}}$	0.23^{***}	0.23^{***}	0.08^{***}	1.00						
		E_{JC}	$\mathbf{0.47^{***}}$	0.07^{**}	$\mathbf{0.47^{***}}$	$\mathbf{0.47^{***}}$	1.00^{***}	0.07^{***}	1.00					
		E_{CCos}	0.20^{***}	0.16^{***}	0.20^{***}	0.20^{***}	0.12^{***}	0.18^{***}	0.12^{***}	1.00				
		RE_C	$\mathbf{0.43^{***}}$	0.19^{***}	$\mathbf{0.43^{***}}$	$\mathbf{0.43^{***}}$	0.35^{***}	0.21^{***}	0.35^{***}	0.23^{***}	1.00			
		RE_T	0.17^{***}	$\mathbf{0.59^{***}}$	0.17^{***}	0.17^{***}	0.01	0.70^{***}	0.01	0.18^{***}	0.21^{***}	1.00		
		RE_{JC}	$\mathbf{0.42^{***}}$	0.16^{***}	$\mathbf{0.42^{***}}$	$\mathbf{0.42^{***}}$	0.35^{***}	0.19^{***}	0.35^{***}	0.23^{***}	1.00^{***}	0.18^{***}	1.00	
		RE_O	$\mathbf{0.42^{***}}$	0.16^{***}	$\mathbf{0.42^{***}}$	$\mathbf{0.42^{***}}$	0.35^{***}	0.19^{***}	0.35^{***}	0.23^{***}	1.00^{***}	0.18^{***}	1.00^{***}	1.00

(b) Seller $\longleftarrow$ Buyer

Features			Location-Based Social Network											
			Topological				Homophilic							
			L_{CN}	L_{TN}	L_{JC}	L_{AA}	E_C	E_T	E_{JC}	E_{CCos}	RE_C	RE_T	RE_{JC}	RE_O
Location-Based Social Network	Topological	L_{CN}	1.00											
		L_{TN}	0.23^{***}	1.00										
		L_{JC}	1.00^{***}	0.23^{***}	1.00									
		L_{AA}	1.00^{***}	0.23^{***}	1.00^{***}	1.00								
	Homophilic	E_C	0.39^{***}	0.04^{*}	0.39^{***}	0.39^{***}	1.00							
		E_T	0.23^{***}	$\mathbf{0.85^{***}}$	0.23^{***}	0.23^{***}	0.05^{*}	1.00						
		E_{JC}	0.39^{***}	0.04	0.39^{***}	0.39^{***}	1.00^{***}	0.04^{*}	1.00					
		E_{CCos}	0.16^{***}	0.15^{***}	0.16^{***}	0.16^{***}	0.12^{***}	0.20^{***}	0.12^{***}	1.00				
		RE_C	$\mathbf{0.42^{***}}$	0.16^{***}	$\mathbf{0.42^{***}}$	$\mathbf{0.42^{***}}$	0.36^{***}	0.19^{***}	0.36^{***}	0.22^{***}	1.00			
		RE_T	0.19^{***}	$\mathbf{0.57^{***}}$	0.19^{***}	0.19^{***}	0.01	0.71^{***}	0.01	0.19^{***}	0.20^{***}	1.00		
		RE_{JC}	$\mathbf{0.41^{***}}$	0.14^{***}	$\mathbf{0.41^{***}}$	$\mathbf{0.41^{***}}$	0.36^{***}	0.17^{***}	0.36^{***}	0.22^{***}	1.00^{***}	0.17^{***}	1.00	
		RE_O	$\mathbf{0.41^{***}}$	0.14^{***}	$\mathbf{0.41^{***}}$	$\mathbf{0.41^{***}}$	0.36^{***}	0.17^{***}	0.36^{***}	0.22^{***}	1.00^{***}	0.17^{***}	1.00^{***}	1.00

the Jaccard's coefficient of the events $E_{JC}(u, v)$ of two users u and v and in the first case also the Jaccard's coefficient of the neighbors in the network $L_{JC}(u, v)$ had values > 0.01. The best AUC value of different constellations of the features was only 53.40%. However, the unsupervised learning results with collaborative filtering were reasonable with partly similar success rates as the online social network features.

In summary, it could be stated that both the topological and the homophilic features of the location-based social network are not that suitable for predicting trading interactions.

Table 6.9: Supervised learning including the AUC values and the information gains and unsupervised learning with collaborative filtering for the location-based social network features. The best information gain and the best success rates were highlighted.

(a) Seller ⟶ Buyer

			Supervised		Unsupervised		
		Features	AUC (LR)	InfoGain	SR@$k=1$	SR@$k=3$	SR@all
Location-Based Social Network	Topological	$L_{CN}(u,v)$	50.90%	< 0.01	0.4497	0.1723	**0.0397**
		$L_{TN}(u,v)$	49.00%	< 0.01	**0.9070**	**0.2839**	0.0194
		$L_{JC}(u,v)$	51.00%	**0.01007**	0.4483	0.1737	**0.0396**
		$L_{AA}(u,v)$	51.00%	< 0.01	0.4483	0.1736	**0.0433**
		Topological	50.30%	—	0.8995	**0.3046**	**0.0421**
	Homophilic	$E_C(u,v)$	50.60%	< 0.01	0.3861	0.1600	**0.0574**
		$E_T(u,v)$	49.50%	< 0.01	**0.9040**	**0.2845**	0.0182
		$E_{JC}(u,v)$	50.60%	**0.01160**	0.3861	0.1594	**0.0558**
		$E_{CCos}(u,v)$	51.10%	< 0.01	0.8160	**0.2966**	0.0086
		$RE_C(u,v)$	51.10%	< 0.01	0.5998	0.2578	**0.0452**
		$RE_T(u,v)$	48.00%	< 0.01	**0.9139**	**0.2855**	0.0131
		$RE_{JC}(u,v)$	51.60%	< 0.01	0.5996	0.2558	**0.0509**
		$RE_O(u,v)$	51.50%	< 0.01	0.5996	0.2558	**0.0509**
		Homophilic	53.20%	—	**0.9025**	**0.3425**	**0.0530**
Location-Based Social Network			53.00%	—	**0.9026**	**0.3359**	**0.0541**

(b) Seller ⟵ Buyer

			Supervised		Unsupervised		
		Features	AUC (LR)	InfoGain	SR@$k=1$	SR@$k=3$	SR@all
Location-Based Social Network	Topological	$L_{CN}(u,v)$	49.90%	< 0.01	0.3432	0.1427	**0.0351**
		$L_{TN}(u,v)$	51.40%	< 0.01	**0.9895**	**0.3472**	0.0077
		$L_{JC}(u,v)$	50.20%	< 0.01	0.3432	0.1443	**0.0367**
		$L_{AA}(u,v)$	49.90%	< 0.01	0.3432	0.1438	**0.0356**
		Topological	53.40%	—	**0.9871**	**0.3749**	**0.0339**
	Homophilic	$E_C(u,v)$	50.60%	< 0.01	0.2953	0.0981	**0.0396**
		$E_T(u,v)$	51.50%	< 0.01	**0.9862**	**0.3462**	0.0088
		$E_{JC}(u,v)$	50.60%	**0.01109**	0.2953	0.0983	**0.0394**
		$E_{CCos}(u,v)$	49.60%	< 0.01	0.8879	**0.3811**	0.0140
		$RE_C(u,v)$	51.10%	< 0.01	0.4862	0.2644	**0.0384**
		$RE_T(u,v)$	49.90%	< 0.01	**0.9902**	**0.3663**	0.0051
		$RE_{JC}(u,v)$	51.30%	< 0.01	0.4868	**0.2730**	**0.0442**
		$RE_O(u,v)$	51.30%	< 0.01	0.4868	**0.2730**	**0.0442**
		Homophilic	52.70%	—	**0.9883**	**0.4079**	**0.0442**
Location-Based Social Network			51.80%	—	**0.9887**	**0.4074**	**0.0429**

6.4 Best data sources and feature sets

This section gives an overview of how the feature sets of all used data sources performed predicting trading interactions. Each feature set on its own was analyzed. A few of the overall 50 used features were prominent and obtained mentionable boosts. The best features for predicting trading interactions were the outgoing preferential attachment score feature of the trading network with AUC values up to 36.20% over the baseline followed by the total neighbors features with 30% over the baseline. These features belong to the topological feature set of the trading network, which achieved the best AUC values up to 35.40% over the baseline. Acceptable predicting values were also perceived with the cosine similarity of product ratings feature with over 21%. This feature was the decisive factor that the homophilic feature set of the trading network reached AUC values up to 23.20%.

Not so well by far, the features and feature sets of the online and location-based social network could not exceed the border of 10% over the baseline both single and combined. Considering only these two networks the best feature was the preferential attachment score of the online social network with the poor value of 6.60% and therefore the best feature set was the topological feature set of the online social network with AUC values up to 9.90% over the baseline.

Table 6.10 shows the experiment results of supervised learning including the AUC values and unsupervised learning with collaborative filtering of all feature sets of all three data sources.

Table 6.10: Supervised learning including the AUC values and unsupervised learning with collaborative filtering for all feature sets of all three data sources.

(a) Seller ⟶ Buyer

Feature Sets		Supervised AUC (LR)	Unsupervised SR@$k=1$	SR@$k=3$	SR@all
Trading	Topological	83.60%	0.9415	0.3434	0.0338
	Homophilic	73.20%	0.8207	0.1839	0.0000
	Overall	84.00%	0.8422	0.2331	0.0244
Online	Topological	59.90%	0.9058	0.3371	0.0293
	Homophilic	57.10%	0.9136	0.3228	0.0197
	Overall	59.80%	0.9084	0.3225	0.0220
Location-Based	Topological	50.30%	0.8995	0.3046	0.0421
	Homophilic	53.20%	0.9025	0.3425	0.0530
	Overall	53.00%	0.9026	0.3359	0.0541

(b) Seller ⟵ Buyer

Feature Sets		Supervised AUC (LR)	Unsupervised SR@$k=1$	SR@$k=3$	SR@all
Trading	Topological	85.40%	0.9967	0.4720	0.0683
	Homophilic	71.80%	0.9849	0.2184	0.0002
	Overall	84.90%	0.9893	0.2991	0.0371
Online	Topological	58.00%	0.9917	0.4144	0.0276
	Homophilic	58.00%	0.9936	0.4160	0.0302
	Overall	59.30%	0.9926	0.4198	0.0316
Location-Based	Topological	53.40%	0.9871	0.3749	0.0339
	Homophilic	52.70%	0.9883	0.4079	0.0442
	Overall	51.80%	0.9887	0.4074	0.0429

6.5 Online and Location-Based Social vs. Trading Network

This section discusses the results of the combination of the online and the location-based social network features to see if any increase could be recognized or even an approximation to the results of the trading network – discussed in Section 6.1 – could be attained.

The correlation matrices in Table 6.11 illustrate the correlations between the online and location-based social network features. The highest significant correlations were observed between the favored regions features $RF_C(u, v)$, $RF_{JC}(u, v)$, $RF_O(u, v)$ for two users u and v and the number of common

events and the Jaccard's coefficient of the events $E_C(u, v)$, $E_{JC}(u, v)$.

The combination of online and location-based social network features had only minor effects for the performance to predict trading interactions. The results in Table 6.12 are almost the same as for online social network features only (see Table 6.6). Consequently location-based social network features were not that suitable to make a trading interaction prediction. Online social network features already operated quite well with probabilities of almost 60%. Consulting the location-based social network features did not result in any increase and could not get anywhere near the results of the prediction with trading network features.

Table 6.11: Spearman's correlation matrices of the online and the location-based social network features with their significance. The best correlations (> 0.20) between these two feature sets were highlighted.

(a) Seller $\longrightarrow$ Buyer

Features			Location-Based Social Network											
			Topological				Homophilic							
			L_{CN}	L_{TN}	L_{JC}	L_{AA}	E_C	E_T	E_{JC}	E_{CCos}	RE_C	RE_T	RE_{JC}	RE_O
Online Social Network	Topological	O^+_{CN}	0.13^{***}	0.05^{*}	0.13^{***}	0.13^{***}	0.19^{***}	0.03	0.19^{***}	0.07^{**}	0.10^{***}	0.04	0.10^{***}	0.10^{***}
		O^-_{CN}	0.12^{***}	0.00	0.12^{***}	0.12^{***}	$\mathbf{0.32^{***}}$	-0.00	$\mathbf{0.32^{***}}$	0.05^{*}	0.10^{***}	-0.01	0.10^{***}	0.10^{***}
		O^+_{TN}	-0.02	0.03	-0.02	-0.02	-0.00	0.03	-0.00	0.01	0.02	0.08^{***}	0.02	0.02
		O^-_{TN}	-0.02	-0.00	-0.02	-0.02	0.01	-0.02	0.01	0.02	-0.03	-0.02	-0.02	-0.02
		O^+_{JC}	0.13^{***}	0.05^{*}	0.13^{***}	0.13^{***}	0.19^{***}	0.03	0.18^{***}	0.07^{**}	0.10^{***}	0.04	0.10^{***}	0.10^{***}
		O^-_{JC}	0.12^{***}	0.00	0.12^{***}	0.12^{***}	$\mathbf{0.32^{***}}$	-0.00	$\mathbf{0.32^{***}}$	0.05^{*}	0.10^{***}	-0.01	0.10^{***}	0.10^{***}
		O^+_{PS}	0.03	0.03	0.03	0.03	0.04^{*}	0.01	0.04^{*}	-0.03	-0.01	0.04^{*}	-0.01	-0.01
		O^-_{PS}	-0.02	0.01	-0.02	-0.02	0.01	0.02	0.01	0.03	0.03	0.03	0.03	0.03
		O_R	0.15^{***}	0.06^{**}	0.15^{***}	0.15^{***}	$\mathbf{0.29^{***}}$	0.06^{**}	$\mathbf{0.29^{***}}$	0.04^{*}	0.11^{***}	0.05^{*}	0.11^{***}	0.11^{***}
		O_{AA}	0.12^{***}	0.01	0.12^{***}	0.12^{***}	$\mathbf{0.28^{***}}$	0.00	$\mathbf{0.28^{***}}$	0.04^{*}	0.09^{***}	-0.00	0.09^{***}	0.09^{***}
	Homophilic	G_C	0.10^{***}	0.01	0.11^{***}	0.11^{***}	0.18^{***}	-0.00	0.18^{***}	0.05^{*}	0.09^{***}	-0.03	0.10^{***}	0.10^{***}
		G_T	0.00	-0.01	0.00	0.00	0.01	-0.01	0.01	-0.01	-0.04^{*}	-0.03	-0.04^{*}	-0.04^{*}
		G_{JC}	0.11^{***}	0.01	0.11^{***}	0.11^{***}	0.18^{***}	0.00	0.18^{***}	0.05^{*}	0.10^{***}	-0.03	0.10^{***}	0.10^{***}
		I_C	0.01	-0.01	0.00	0.01	0.02	-0.02	0.02	0.01	0.02	-0.01	0.02	0.02
		I_T	-0.02	-0.01	-0.02	-0.02	-0.02	0.01	-0.02	-0.05^{*}	0.04^{*}	0.02	0.04^{*}	0.04^{*}
		I_{JC}	0.01	-0.01	0.00	0.01	0.02	-0.02	0.02	0.01	0.02	-0.01	0.02	0.02
		OI	0.06^{**}	0.03	0.06^{**}	0.06^{**}	0.19^{***}	0.02	0.19^{***}	0.04^{*}	0.08^{***}	0.00	0.08^{***}	0.08^{***}
		RR_C	0.05^{*}	0.04^{*}	0.05^{*}	0.05^{*}	0.10^{***}	0.03	0.10^{***}	0.04^{*}	0.02	0.04^{*}	0.02	0.02
		RR_T	0.00	0.02	0.00	0.00	0.01	0.02	0.01	-0.04^{*}	-0.02	0.02	-0.02	-0.02
		RR_{JC}	0.05^{*}	0.04^{*}	0.05^{*}	0.05^{*}	0.10^{***}	0.03	0.10^{***}	0.04^{*}	0.02	0.04^{*}	0.02	0.02
		RR_O	0.05^{*}	0.04^{*}	0.05^{*}	0.05^{*}	0.10^{***}	0.03	0.10^{***}	0.04^{*}	0.02	0.04^{*}	0.02	0.02
		RF_C	0.18^{***}	-0.02	0.18^{***}	0.18^{***}	$\mathbf{0.41^{***}}$	-0.01	$\mathbf{0.41^{***}}$	0.02	0.17^{***}	-0.02	0.18^{***}	0.18^{***}
		RF_T	0.02	0.03	0.02	0.02	-0.01	0.07^{**}	-0.01	-0.00	0.02	0.09^{***}	0.02	0.02
		RF_{JC}	0.18^{***}	-0.02	0.18^{***}	0.18^{***}	$\mathbf{0.41^{***}}$	-0.01	$\mathbf{0.41^{***}}$	0.02	0.17^{***}	-0.02	0.18^{***}	0.18^{***}
		RF_O	0.18^{***}	-0.02	0.18^{***}	0.18^{***}	$\mathbf{0.41^{***}}$	-0.01	$\mathbf{0.41^{***}}$	0.02	0.17^{***}	-0.02	0.18^{***}	0.18^{***}

(b) Seller $\longleftarrow$ Buyer

Features			Location-Based Social Network											
			Topological				Homophilic							
			L_{CN}	L_{TN}	L_{JC}	L_{AA}	E_C	E_T	E_{JC}	E_{CCos}	RE_C	RE_T	RE_{JC}	RE_O
Online Social Network	Topological	O^+_{CN}	0.13^{***}	0.03	0.13^{***}	0.13^{***}	0.09^{***}	0.02	0.09^{***}	0.05^{*}	0.08^{***}	0.03	0.08^{***}	0.08^{***}
		O^-_{CN}	0.10^{***}	0.00	0.10^{***}	0.10^{***}	$\mathbf{0.23^{***}}$	0.00	$\mathbf{0.24^{***}}$	0.05^{*}	0.11^{***}	0.02	0.11^{***}	0.11^{***}
		O^+_{TN}	0.04^{*}	0.05^{*}	0.04^{*}	0.04^{*}	0.01	0.07^{**}	0.01	0.02	0.04	0.11^{***}	0.03	0.03
		O^-_{TN}	0.05^{*}	0.02	0.05^{*}	0.05^{*}	-0.02	0.01	-0.02	0.05^{*}	-0.02	0.03	-0.03	-0.03
		O^+_{JC}	0.13^{***}	0.03	0.13^{***}	0.13^{***}	0.09^{***}	0.02	0.09^{***}	0.05^{*}	0.08^{***}	0.03	0.08^{***}	0.08^{***}
		O^-_{JC}	0.10^{***}	0.00	0.10^{***}	0.10^{***}	$\mathbf{0.23^{***}}$	0.00	$\mathbf{0.24^{***}}$	0.05^{*}	0.11^{***}	0.02	0.11^{***}	0.11^{***}
		O^+_{PS}	0.02	-0.01	0.02	0.02	-0.02	0.01	-0.02	0.04	-0.03	0.04^{*}	-0.03	-0.03
		O^-_{PS}	0.04^{*}	0.05^{*}	0.04^{*}	0.04^{*}	0.05^{*}	0.05^{*}	0.05^{*}	0.01	0.05^{*}	0.09^{***}	0.05^{*}	0.05^{*}
		O_R	0.09^{***}	0.01	0.10^{***}	0.10^{***}	$\mathbf{0.24^{***}}$	0.01	$\mathbf{0.24^{***}}$	0.05^{*}	0.12^{***}	-0.00	0.12^{***}	0.12^{***}
		O_{AA}	0.12^{***}	0.04^{*}	0.12^{***}	0.12^{***}	0.13^{***}	0.02	0.13^{***}	0.05^{*}	0.08^{***}	0.04	0.07^{***}	0.07^{***}
	Homophilic	G_C	0.06^{**}	-0.04^{*}	0.06^{**}	0.06^{**}	0.13^{***}	-0.04^{*}	0.13^{***}	0.05^{*}	0.10^{***}	-0.04^{*}	0.11^{***}	0.11^{***}
		G_T	0.00	0.02	0.00	0.00	-0.05^{*}	0.02	-0.05^{*}	-0.03	-0.03	0.02	-0.03	-0.03
		G_{JC}	0.06^{**}	-0.04^{*}	0.06^{**}	0.06^{**}	0.13^{***}	-0.04^{*}	0.13^{***}	0.05^{*}	0.10^{***}	-0.05^{*}	0.11^{***}	0.11^{***}
		I_C	0.05^{*}	0.03	0.05^{*}	0.05^{*}	0.02	0.03	0.02	0.03	0.04^{*}	0.03	0.04^{*}	0.04^{*}
		I_T	0.02	0.03	0.01	0.02	0.01	0.03	0.01	-0.00	0.02	0.05^{*}	0.03	0.03
		I_{JC}	0.05^{*}	0.03	0.05^{*}	0.05^{*}	0.02	0.03	0.02	0.03	0.04^{*}	0.03	0.04^{*}	0.04^{*}
		OI	0.07^{**}	0.00	0.07^{**}	0.07^{**}	0.19^{***}	0.01	0.19^{***}	0.03	0.08^{***}	0.02	0.08^{***}	0.08^{***}
		RR_C	0.08^{***}	0.04	0.08^{***}	0.08^{***}	0.07^{**}	0.03	0.07^{**}	0.05^{*}	0.04^{*}	0.03	0.04^{*}	0.04^{*}
		RR_T	0.02	0.02	0.02	0.02	0.01	0.00	0.01	-0.05^{*}	0.02	0.03	0.02	0.02
		RR_{JC}	0.08^{***}	0.04	0.08^{***}	0.08^{***}	0.07^{**}	0.03	0.07^{**}	0.05^{*}	0.04^{*}	0.03	0.04^{*}	0.04^{*}
		RR_O	0.08^{***}	0.04	0.08^{***}	0.08^{***}	0.07^{**}	0.03	0.07^{**}	0.05^{*}	0.04^{*}	0.03	0.04^{*}	0.04^{*}
		RF_C	0.13^{***}	-0.03	0.13^{***}	0.13^{***}	$\mathbf{0.45^{***}}$	-0.01	$\mathbf{0.45^{***}}$	0.07^{**}	$\mathbf{0.21^{***}}$	-0.04^{*}	$\mathbf{0.22^{***}}$	$\mathbf{0.22^{***}}$
		RF_T	0.03	0.12^{***}	0.03	0.03	-0.03	0.13^{***}	-0.03	0.02	0.05^{*}	0.14^{***}	0.05^{*}	0.05^{*}
		RF_{JC}	0.13^{***}	-0.03	0.13^{***}	0.13^{***}	$\mathbf{0.45^{***}}$	-0.01	$\mathbf{0.45^{***}}$	0.07^{**}	$\mathbf{0.21^{***}}$	-0.04^{*}	$\mathbf{0.22^{***}}$	$\mathbf{0.22^{***}}$
		RF_O	0.13^{***}	-0.03	0.13^{***}	0.13^{***}	$\mathbf{0.45^{***}}$	-0.01	$\mathbf{0.45^{***}}$	0.07^{**}	$\mathbf{0.21^{***}}$	-0.04^{*}	$\mathbf{0.22^{***}}$	$\mathbf{0.22^{***}}$

Table 6.12: Supervised learning including the AUC values and unsupervised learning with collaborative filtering for the combination of the online and the location-based social network features in comparison with the trading network features.

(a) Seller $\longrightarrow$ Buyer

Feature Sets		Supervised AUC (LR)	Unsupervised SR@$k = 1$	SR@$k = 3$	SR@all
Trading	Topological	83.60%	0.9415	0.3434	0.0338
	Homophilic	73.20%	0.8207	0.1839	0.0000
	Overall	84.00%	0.8422	0.2331	0.0244
Online + Location-Based	Topological	58.60%	0.8959	0.3170	0.0415
	Homophilic	57.40%	0.9113	0.3287	0.0264
	Overall	59.90%	0.9125	0.3279	0.0364

(b) Seller $\longleftarrow$ Buyer

Feature Sets		Supervised AUC (LR)	Unsupervised SR@$k = 1$	SR@$k = 3$	SR@all
Trading	Topological	85.40%	0.9967	0.4720	0.0683
	Homophilic	71.80%	0.9849	0.2184	0.0002
	Overall	84.90%	0.9893	0.2991	0.0371
Online + Location-Based	Topological	57.80%	0.9870	0.3846	0.0422
	Homophilic	58.30%	0.9908	0.4148	0.0344
	Overall	59.50%	0.9898	0.4071	0.0388

6.6 Combination of Network Features

The previous sections indicated that the features of the trading network were the best for the prediction of trading interactions by far. This section describes the results of the experiments of several combinations of the trading network with the online and/or the location-based social network to show if the addition of further features to the trading network resulted in a considerable increase.

The feature analysis in terms of the correlation matrices in Table 6.13 states high correlations between the reciprocity of the user communication feature of the trading network $T_R(u, v)$ and the favored regions features and the interaction feature of the online social network $RF_C(u, v)$, $RF_{JC}(u, v)$, $RF_O(u, v)$, $OI(u, v)$ on the one hand and the number of common events and the number of total events of the location-based social network $E_C(u, v)$, $E_T(u, v)$ on the other hand.

The results in Table 6.14 for the combination of the features of the different networks illustrate that the addition of the online or the location-based social network features or both to the trading network feature did not result in a recognizable effect. The AUC values and success rates are almost identical with the values in Table 6.3 showing the results for the trading interaction prediction considering only the trading network features.

Table 6.13: Spearman's correlation matrices between the trading network features and the online and location-based social network features with their significance. The best correlations (> 0.20) between these feature sets were highlighted.

(a) Seller $\longrightarrow$ Buyer

Features			Trading Network												
			Topological									Homophilic			
			T^{+}_{CN}	T^{-}_{CN}	T^{+}_{TN}	T^{-}_{TN}	T^{+}_{JC}	T^{-}_{JC}	T^{+}_{PS}	T^{-}_{PS}	T_{R}	T_{AA}	P_{CCos}	P_{PCos}	P_{RCos}
Online Social Network	Topological	O^{+}_{CN}	-0.01	0.00	-0.02	-0.01	-0.01	0.00	-0.03	-0.02	0.09***	-0.01	-0.01	-0.07**	0.03
		O^{-}_{CN}	0.04	0.04*	-0.01	-0.00	0.04	0.04*	-0.02	0.01	**0.21*** **	0.08***	-0.00	0.01	0.04*
		O^{+}_{TN}	-0.05*	0.02	-0.08***	0.07**	-0.05*	0.02	-0.03	0.00	0.01	0.01	-0.04*	-0.03	0.00
		O^{-}_{TN}	-0.00	-0.03	0.16***	0.03	-0.00	-0.03	0.11***	0.05*	0.01	-0.00	0.03	0.04*	0.00
		O^{+}_{JC}	-0.01	0.00	-0.02	-0.01	-0.01	0.00	-0.03	-0.02	0.09***	-0.01	-0.01	-0.07**	0.02
		O^{-}_{JC}	0.04	0.04*	-0.01	-0.00	0.04	0.04*	-0.02	0.01	**0.22*** **	0.08***	-0.00	0.01	0.04*
		O^{+}_{PS}	-0.01	0.02	-0.14***	-0.02	-0.01	0.02	-0.18***	0.09***	0.03	0.00	-0.11***	-0.08***	-0.03
		O^{-}_{PS}	-0.07**	-0.02	0.12***	0.08***	-0.07**	-0.02	0.15***	-0.04*	-0.02	-0.00	0.07**	0.05*	0.03
		O_{R}	-0.01	0.02	-0.01	0.01	-0.01	0.02	-0.03	-0.00	0.13***	-0.01	-0.01	-0.02	0.04*
		O_{AA}	0.04*	0.01	-0.00	-0.00	0.04*	0.01	-0.02	0.01	0.22***	0.04*	-0.01	0.01	0.05*
	Homophilic	G_{C}	0.02	0.04*	0.03	0.01	0.02	0.04*	0.01	0.05*	0.16***	0.03	0.06*	0.03	0.05*
		G_{T}	-0.04*	-0.02	-0.06**	-0.08***	-0.04	-0.02	-0.11***	0.03	0.03	-0.03	-0.10***	-0.05*	-0.04*
		G_{JC}	0.02	0.04*	0.03	0.01	0.02	0.04*	0.01	0.05*	0.16***	0.03	0.06**	0.03	0.05*
		I_{C}	-0.02	0.05*	0.01	0.05*	-0.02	0.05*	0.01	0.08***	-0.01	0.04*	-0.00	0.01	-0.02
		I_{T}	0.03	0.05*	0.00	0.15***	0.03	0.04*	0.07**	0.09***	-0.01	0.00	0.02	0.03	-0.01
		I_{JC}	-0.02	0.05*	0.01	0.05*	-0.02	0.05*	0.01	0.08***	-0.01	0.04*	-0.00	0.01	-0.02
		OI	-0.01	0.07**	-0.00	0.02	-0.01	0.07**	-0.02	0.06**	**0.31*** **	0.06*	0.00	-0.01	0.05*
		RR_{C}	-0.01	-0.01	0.00	-0.02	-0.01	-0.01	-0.01	-0.01	-0.00	-0.01	-0.03	0.02	0.02
		RR_{T}	-0.02	0.02	-0.05*	0.05*	-0.02	0.02	-0.03	0.05*	0.03	-0.01	-0.03	-0.04*	-0.04*
		RR_{JC}	-0.01	-0.01	0.00	-0.02	-0.01	-0.01	-0.01	-0.01	-0.00	-0.01	-0.03	0.02	0.02
		RR_{O}	-0.01	-0.01	0.00	-0.02	-0.01	-0.01	-0.01	-0.01	-0.00	-0.01	-0.03	0.02	0.02
		RF_{C}	0.12***	0.07***	0.01	-0.01	0.12***	0.08***	-0.03	0.07**	**0.30*** **	0.05*	0.08***	0.04*	0.03
		RF_{T}	-0.04*	0.04*	0.00	0.03	-0.04*	0.04*	0.01	0.04	0.02	-0.01	0.04*	-0.00	0.04*
		RF_{JC}	0.11***	0.08***	0.01	-0.01	0.11***	0.08***	-0.03	0.07**	**0.30*** **	0.05*	0.08***	0.04*	0.03
		RF_{O}	0.11***	0.08***	0.01	-0.01	0.11***	0.08***	-0.03	0.07**	**0.30*** **	0.05*	0.08***	0.04*	0.03
Location-Based Social Network	Topological	L_{CN}	0.05*	0.02	-0.05*	0.02	0.05*	0.02	-0.03	-0.01	0.15***	0.02	0.05*	0.01	0.04*
		L_{TN}	-0.05*	-0.02	-0.07***	0.02	-0.05*	-0.02	-0.04*	-0.06*	0.01	0.01	0.04*	0.04	0.04*
		L_{JC}	0.06*	0.02	-0.05*	0.02	0.06*	0.02	-0.03	-0.01	0.15***	0.02	0.05*	0.01	0.04*
		L_{AA}	0.05*	0.02	-0.05*	0.02	0.05*	0.02	-0.03	-0.01	0.15***	0.02	0.05*	0.01	0.04*
	Homophilic	E_{C}	0.10***	0.05*	-0.06**	0.01	0.11***	0.05*	-0.08***	0.03	**0.35*** **	0.07**	0.04*	0.03	0.05*
		E_{T}	-0.04*	-0.00	-0.05*	0.04*	-0.04*	-0.00	-0.01	-0.05*	0.03	0.02	0.07**	0.04*	0.05*
		E_{JC}	0.11***	0.05*	-0.06**	0.01	0.11***	0.05*	-0.08***	0.03	**0.35*** **	0.07**	0.04*	0.03	0.05*
		E_{CCos}	-0.02	0.01	-0.07***	0.03	-0.02	0.01	-0.05*	-0.05*	0.05*	0.06*	0.04*	-0.03	0.03
		RE_{C}	0.03	0.05*	-0.03	0.03	0.03	0.05*	-0.02	0.00	0.12***	0.04	0.06**	-0.00	0.03
		RE_{T}	-0.07**	-0.00	-0.06**	0.04*	-0.07**	-0.00	-0.02	-0.07**	-0.00	0.01	0.05*	0.02	0.04*
		RE_{JC}	0.04	0.04*	-0.02	0.03	0.04	0.05*	-0.02	0.00	0.13***	0.03	0.07**	-0.00	0.03
		RE_{O}	0.04	0.04*	-0.02	0.03	0.04	0.05*	-0.02	0.00	0.13***	0.03	0.07**	-0.00	0.03

(b) Seller $\longleftarrow$ Buyer

Features			Trading Network												
			Topological									Homophilic			
			T^{+}_{CN}	T^{-}_{CN}	T^{+}_{TN}	T^{-}_{TN}	T^{+}_{JC}	T^{-}_{JC}	T^{+}_{PS}	T^{-}_{PS}	T_{R}	T_{AA}	P_{CCos}	P_{PCos}	P_{RCos}
Online Social Network	Topological	O^{+}_{CN}	0.01	-0.01	0.01	-0.02	0.01	-0.01	-0.02	-0.01	-0.01	-0.01	-0.02	-0.01	0.01
		O^{-}_{CN}	0.03	-0.01	-0.01	-0.02	0.03	-0.01	-0.03	0.02	0.14***	-0.01	-0.01	-0.04*	-0.01
		O^{+}_{TN}	-0.04*	-0.03	0.07***	-0.09***	-0.04*	-0.03	-0.04	0.01	-0.03	-0.04*	-0.01	-0.04*	0.02
		O^{-}_{TN}	-0.06**	0.02	0.05*	0.16***	-0.06**	0.02	0.12***	0.08***	0.01	0.01	0.06**	0.01	0.07**
		O^{+}_{JC}	0.01	-0.01	0.01	-0.02	0.01	-0.01	-0.02	-0.01	-0.01	-0.01	-0.02	-0.01	0.01
		O^{-}_{JC}	0.03	-0.01	-0.01	-0.02	0.03	-0.01	-0.03	0.02	0.14***	-0.01	-0.01	-0.04*	-0.01
		O^{+}_{PS}	-0.04	-0.03	0.10***	0.12***	-0.04	-0.03	0.17***	-0.02	-0.04*	-0.03	0.09***	0.06**	0.06**
		O^{-}_{PS}	-0.03	-0.02	-0.00	-0.13***	-0.03	-0.02	-0.15***	0.06*	0.01	-0.02	-0.07***	-0.11***	-0.02
		O_{R}	0.07**	-0.01	0.02	-0.04*	0.07**	-0.01	-0.05*	0.06*	**0.22*** **	-0.00	0.01	-0.04*	0.02
		O_{AA}	0.01	-0.01	-0.01	-0.02	0.01	-0.01	-0.03	-0.01	-0.01	-0.01	-0.02	-0.03	-0.01
	Homophilic	G_{C}	0.03	0.01	0.00	-0.00	0.03	0.01	-0.04*	0.04	0.17***	0.03	0.06*	0.02	0.01
		G_{T}	-0.06*	-0.09***	-0.07***	-0.10***	-0.05*	-0.09***	-0.11***	0.03	-0.01	-0.04	-0.07**	-0.03	-0.07**
		G_{JC}	0.03	0.01	0.00	0.00	0.03	0.01	-0.04*	0.04*	0.18***	0.03	0.06**	0.02	0.01
		I_{C}	0.04	-0.01	0.05*	0.03	0.04	-0.01	0.03	0.04*	-0.01	-0.01	-0.01	0.04*	-0.01
		I_{T}	0.05*	-0.01	0.17***	0.00	0.04*	-0.01	0.06**	0.07**	-0.01	0.00	0.01	0.01	0.06*
		I_{JC}	0.04	-0.01	0.05*	0.03	0.04	-0.01	0.03	0.04*	-0.01	-0.01	-0.01	0.04*	-0.01
		OI	-0.01	-0.01	0.01	-0.04*	-0.01	-0.01	-0.05*	0.01	0.12***	-0.00	-0.02	-0.04*	-0.00
		RR_{C}	-0.01	-0.01	0.00	0.01	-0.01	-0.01	-0.01	0.01	-0.00	-0.00	-0.02	-0.02	0.02
		RR_{T}	-0.03	-0.00	0.03	-0.05*	-0.03	-0.00	-0.04*	0.06**	0.03	-0.01	0.00	-0.03	-0.00
		RR_{JC}	-0.01	-0.01	0.00	0.01	-0.01	-0.01	-0.01	0.01	-0.00	-0.00	-0.02	-0.02	0.02
		RR_{O}	-0.01	-0.01	0.00	0.01	-0.01	-0.01	-0.01	0.01	-0.00	-0.00	-0.02	-0.02	0.02
		RF_{C}	-0.00	0.06**	0.01	-0.04*	-0.00	0.06**	-0.05*	0.05*	**0.32*** **	0.05*	0.02	-0.01	0.01
		RF_{T}	0.02	-0.04*	0.01	-0.08***	0.02	-0.04*	-0.07**	-0.01	0.00	-0.02	0.01	-0.01	0.04*
		RF_{JC}	-0.00	0.06**	0.01	-0.04*	-0.00	0.06**	-0.05*	0.05*	**0.32*** **	0.05*	0.02	-0.01	0.01
		RF_{O}	-0.00	0.06**	0.01	-0.04*	-0.00	0.06**	-0.05*	0.05*	**0.32*** **	0.05*	0.02	-0.01	0.01
Location-Based Social Network	Topological	L_{CN}	0.02	0.01	0.02	-0.07**	0.02	0.01	-0.05*	-0.00	0.02	-0.02	0.05*	0.03	0.03
		L_{TN}	-0.01	-0.02	0.04*	-0.09***	-0.01	-0.02	-0.04*	-0.05*	-0.01	-0.00	0.02	0.03	0.07**
		L_{JC}	0.02	0.01	0.02	-0.07**	0.02	0.01	-0.05*	-0.00	0.02	-0.02	0.05*	0.03	0.03
		L_{AA}	0.02	0.01	0.02	-0.07**	0.02	0.01	-0.05*	-0.00	0.02	-0.02	0.05*	0.03	0.03
	Homophilic	E_{C}	0.05*	0.05*	-0.02	-0.06**	0.05*	0.05*	-0.09***	0.02	0.21***	0.04*	0.05*	0.03	0.01
		E_{T}	0.01	-0.02	0.05*	-0.07**	0.01	-0.02	-0.03	-0.05*	0.01	-0.00	0.03	0.03	0.07**
		E_{JC}	0.05*	0.05*	-0.02	-0.06**	0.05*	0.05*	-0.09***	0.02	0.21***	0.04*	0.05*	0.03	0.01
		E_{CCos}	0.03	-0.00	0.06**	-0.02	0.03	-0.00	0.01	-0.07**	0.02	-0.01	0.02	-0.02	0.05*
		RE_{C}	0.05*	0.00	-0.00	-0.06**	0.05*	0.00	-0.05*	-0.02	0.10***	0.01	0.04*	0.00	0.04
		RE_{T}	0.02	-0.04*	0.03	-0.09***	0.02	-0.04*	-0.06**	-0.07***	-0.02	-0.03	0.02	-0.03	0.04
		RE_{JC}	0.05*	0.01	-0.00	-0.05*	0.05*	0.01	-0.05*	-0.02	0.11***	0.02	0.05*	0.00	0.04*
		RE_{O}	0.05*	0.01	-0.00	-0.05*	0.05*	0.01	-0.05*	-0.02	0.11***	0.02	0.05*	0.00	0.04*

Table 6.14: Supervised learning including the AUC values and unsupervised learning with collaborative filtering for the combination of the trading network features with the online and location-based social network features.

(a) Seller ⟶ Buyer

Feature Sets		Supervised AUC (LR)	Unsupervised SR@$k=1$	SR@$k=3$	SR@all
Trading	Topological	83.60%	0.9415	0.3434	0.0338
	Homophilic	73.20%	0.8207	0.1839	0.0000
	Overall	84.00%	0.8422	0.2331	0.0244
Trading + Online	Topological	83.70%	0.9243	0.3424	0.0431
	Homophilic	73.70%	0.9120	0.3092	0.0190
	Overall	84.50%	0.8958	0.3102	0.0314
Trading + Location-Based	Topological	84.20%	0.9139	0.3292	0.0462
	Homophilic	73.40%	0.8445	0.2398	0.0355
	Overall	84.20%	0.8615	0.2721	0.0486
Trading + Online + Location-Based	Topological	83.80%	0.9107	0.3327	0.0490
	Homophilic	73.60%	0.8810	0.3012	0.0261
	Overall	84.20%	0.9097	0.3170	0.0424

(b) Seller ⟵ Buyer

Feature Sets		Supervised AUC (LR)	Unsupervised SR@$k=1$	SR@$k=3$	SR@all
Trading	Topological	85.40%	0.9967	0.4720	0.0683
	Homophilic	71.80%	0.9849	0.2184	0.0002
	Overall	84.90%	0.9893	0.2991	0.0371
Trading + Online	Topological	85.10%	0.9938	0.4499	0.0721
	Homophilic	72.70%	0.9935	0.3822	0.0217
	Overall	84.80%	0.9932	0.4030	0.0474
Trading + Location-Based	Topological	85.90%	0.9929	0.4406	0.0738
	Homophilic	72.50%	0.9851	0.2891	0.0378
	Overall	85.20%	0.9897	0.3499	0.0668
Trading + Online + Location-Based	Topological	85.60%	0.9927	0.4241	0.0754
	Homophilic	73.20%	0.9888	0.3702	0.0292
	Overall	84.70%	0.9919	0.4068	0.0600

7 Conclusions

This chapter concludes the work. The findings and contributions are briefly highlighted and also an outlook on feasible future work is provided.

7.1 Summary of Findings

Summarizing the experiments, it can be stated that the location-based social network features attained the worst results with AUC values around 53%. The preferential attachment score features $O^+_{PS}(u,v)$, $O^-_{PS}(u,v)$ with AUC values of more than 56% and the total neighbors $T^+_{TN}(u,v)$, $T^-_{TN}(u,v)$, the total groups $G_T(u,v)$ and total favored regions $RF_T(u,v)$ with high success rates with collaborative filtering were the best of the online social network features. The combination of all topological online social network features achieved a prediction probability of nearly 60%. The homophilic features with ≈57% did not reflect a crucial increase in combination with the topological features. Also the combination of all online social network features with all location-based social network features did not result in a considerable improvement in comparison with the results of the topological online social network features only.

To exceed the border of a 60% prediction probability, the trading network features had to be considered. The best feature performance was observed with the outgoing preferential attachment score feature $T^+_{PS}(u,v)$ with AUC values up to 86.20% and also the highest information gain and success rates. Although the total neighbors features $T^+_{TN}(u,v)$, $T^-_{TN}(u,v)$ also reached results of almost 80%, the combination of all topological trading network fea-

tures did not result in an increase in comparison with the results of the outgoing preferential attachment score feature on its own.

Furthermore, the combination of the online and/or the location-based social network feature sets with the trading network features did not significantly improve the results of the trading network features only.

Conclusively, it could be said that the location-based social network features are not suitable for a trading interaction prediction. By contrast, the online social network features provided a passable performance, but could not result in an increase in combination with the dominant trading network features, which revealed very good results.

7.2 Answer to Research Questions

The goal of this master thesis was to show to what extent trading interactions between user pairs can be predicted based on different sources. The research questions of this thesis (see Section 1.2) could briefly be answered as follows:

- To answer the first three research questions, data from three Second Life related sources were collected – an online social network, a location-based social network and a trading network. The data were analyzed and prepared for the computation of topological and homophilic features and the usage in the experiments. As expected, the best trading prediction results in the experiments were obtained using information of the trading network. AUC values up to 35.90% over the baseline could be achieved. To know online social network data could implicate passable prediction results up to ≈10% over the baseline. Disappointingly, having only location-based social information about the users will not result in a crucial success rate regarding the prediction of trading interactions.

- To answer the fourth research question, each feature set was separately analyzed. Since the outgoing preferential attachment score feature and the total neighbors features of the trading network as the best performing features belong to the topological feature set of the trading network, this feature set was the best by far with AUC values up to 35.40% over the baseline. Also an acceptable predicting performance was perceived with the homophilic feature set of the trading network up to 23.20%.

Not half as good, the features and feature sets of the online and location-based social network could not exceed the border of 10% over the baseline both single and combined. Considering only these two networks the best feature set was the topological feature set of the online social network with AUC values up to 9.90% over the baseline.

- For the fifth research question, the combination of online and location-based social network features was necessary. Unfortunately, no recognizable effect for the performance to predict trading interactions was detected. The success rates achieved with the dominant trading network features could not be reached.

- Although, the sixth research question is about the combination of all three network features, the achieved AUC values were almost identically with the values for the trading interaction prediction considering only the trading network features. In conclusion, online social network information could result in passable performance, but knowing trading network information does not require the addition of further information of other network sources for trading interaction predictions.

In summary, it could be stated that for powerful trading relation predictions there is no way around using attributes originating from the trading network, not even having a vast number of combined online and location-based social network features.

7.3 Limitations and Future Directions

As already mentioned, this thesis is focused on predicting trading interactions based on features of several network sources. In the future the time component could be a very interesting factor, which was entirely neglected in this work. Time-dependent attributes could be used as prediction features or existing features could be adapted to refine the trading prediction results. For example, when calculating a feature between two users that is about the already traded products, the products could be weighted in a way where the older trades would not be that crucial as newer ones.

Furthermore, the user coverage for the different feature attributes could be included to improve the collaborative filtering results. This means that for

a feature calculation only those users are considered, for whom information about the current attribute is available.

Item to user – or in this case product to buyer – recommendations based on the existing data could also be an interesting point for future work.

Since with the data of Second Life the experiments in this thesis were based on a virtual world, an important part in the future could be to investigate how the experiments would perform using data of the "real" world.

Finally, since using the data of Second Life, the experiments in this thesis were based on a virtual world. An important task in the future could be to investigate how the experiments would perform if data of the "real" world were used.

Bibliography

[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003. 6, 35

[2] L. Backstrom and J. Kleinberg. Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook. *ArXiv e-prints*, 2013. 5

[3] A.-L. Barabasi and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, 1999. 7, 35

[4] A.-L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311:590–614, 2002. 7

[5] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When Is "Nearest Neighbor" Meaningful? In *Proceedings of the 7th International Conference on Database Theory*, pages 217–235. Springer-Verlag, 1999. 48

[6] K. Bischoff. We love rock 'n' roll: analyzing and predicting friendship links in Last.fm. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 47–56. ACM, 2012. 11

[7] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. 8

[8] S. Boslaugh. *Statistics in a nutshell.* O'Reilly Media, Inc., 2012. 28

[9] J. Cheng, D. M. Romero, B. Meeder, and J. Kleinberg. Predicting Reciprocity in Social Networks. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 49–56, 2011. 10, 35

[10] J. S. Coleman. Social Capital in the Creation of Human Capital. *American Journal of Sociology*, pages 95–120, 1988. 1, 6

[11] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 119–128. ACM, 2010. 9, 34, 36, 38

[12] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, SIGCOMM '99, pages 251–262. ACM, 1999. 31

[13] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici. Link Prediction in Social Networks using Computationally Efficient Topological Features. In *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust (PASSAT) / 2011 IEEE Third Int'l Conference on Social Computing (SocialCom)*, pages 73–80. IEEE, 2011. 9, 10

[14] M. Fire, L. Tenenboim-Chekina, R. Puzis, O. Lesser, L. Rokach, and Y. Elovici. Computationally Efficient Link Prediction in a Variety of Social Networks. *ACM TIST*, 5(1):10:1–10:25, 2013. 7, 10

[15] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of Trust and Distrust. In *Proceedings of the 13th International Conference on World Wide Web*, pages 403–412. ACM, 2004. 47

[16] S. Guo, M. Wang, and J. Leskovec. The role of social networks in online shopping: information passing, price of trust, and consumer choice. In *ACM Conference on Electronic Commerce*, pages 157–166, 2011. 1, 5, 6

[17] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *Proceedings of SDM 06 workshop on Link Analysis, Counterterrorism and Security*, 2006. 9, 10

[18] J. J. Jones, J. E. Settle, R. M. Bond, C. J. Fariss, C. Marlow, and J. H. Fowler. Inferring Tie Strength from Online Directed Behavior. *PLoS ONE*, 8(1), 2013. 10

[19] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953. 7

[20] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. GroupLens: Applying Collaborative Filtering to Usenet News. *Communications of the ACM*, 40(3):77–87, 1997. 11

[21] V. Krebs. Mapping networks of terrorist cells. *CONNECTIONS*, 24(3):43–52, 2002. 5

[22] C.-A. La and P. Michiardi. Characterizing User Mobility in Second Life. In *Proceedings of the first workshop on Online social networks*, WOSN '08. ACM, 2008. 16

[23] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting Positive and Negative Links in Online Social Networks. In *Proceedings of the 19th International Conference on World Wide Web*, pages 641–650. ACM, 2010. 10

[24] H. Levene. *In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I. Olkin et al. eds*, pages 278–292. Stanford University Press, 1960. 49

[25] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007. 1, 5, 6, 48

[26] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Riew of Sociology*, 27(1):415–444, 2001. 8

[27] T. Murata and S. Moriyasu. Link Prediction of Social Networks Based on Weighted Proximity Measures. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 85–88, 2007. 10

[28] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102, 2001. 6, 7

[29] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, pages 175–186. ACM, 1994. 11

[30] M. Rowe, M. Stankovic, and H. Alani. Who will follow whom? Exploiting Semantics for Link Prediction in Attention-Information Networks. In *Proceedings of the 11th international conference on The Semantic Web - Volume Part I*, ISWC'12, pages 476–491. Springer-Verlag, 2012. 10, 33

[31] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1983. 6

[32] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based Collaborative Filtering Recommendation Algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, pages 285–295. ACM, 2001. 11

[33] M. Steurer and C. Trattner. Acquaintance or Partner? Predicting Partnership in Online and Location-Based Social Networks. In *IEEE/ACM ASONAM*, pages 1–8, 2013. 6, 7, 8, 9, 10, 11, 13, 17, 34, 35, 38, 41, 47

[34] M. Steurer and C. Trattner. Predicting Interactions in Online Social Networks: An Experiment in Second Life. In *Proceedings of the 4th International Workshop on Modeling Social Media*, pages 5:1–5:8. ACM, 2013. 10, 13, 33

[35] M. Steurer and C. Trattner. Who will Interact with Whom? A Case-Study in Second Life Using Online Social Network and Location-Based Social Network Features to Predict Interactions between Users. In *Ubiquitous Social Media Analysis*, volume 8329 of *Lecture Notes in Computer Science*, pages 108–127. Springer Berlin Heidelberg, 2013. 10, 13

[36] M. Steurer, C. Trattner, and D. Helic. Predicting Social Interactions from Different Sources of Location-based Knowledge. In *The Third International Conference on Social Eco-Informatics*, SOTICS 2013, pages 8–13, Lisbon, Portugal, 2013. 10, 13, 35, 36

[37] M. Steurer, C. Trattner, and F. Kappe. Success factors of events in virtual worlds a case study in second life. In *Proceedings of the 11th Annual Workshop on Network and Systems Support for Games*, NetGames '12, pages 19:1–19:2. IEEE Press, 2012. 23

[38] M. Thelwall. Homophily in MySpace. *Journal of the American Society for Information Science and Technology*, 60(2):219–231, 2009. 8, 9

[39] C. Trattner, D. Parra, L. Eberhard, and X. Wen. Who will Trade with Whom? Predicting Buyer-Seller Interactions in Online Trading Platforms through Social Networks. In *WWW'14 Companion*, Seoul, Korea, 2014. 10

[40] M. Varvello, F. Picconi, C. Diot, and E. Biersack. Is There Life in Second Life? In *Conext'08*, Madrid, Spain, 2008. 14, 15, 16

[41] M. Varvello and G. M. Voelker. Second Life: A Social Network of Humans and Bots. In *Proceedings of the 20th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, NOSSDAV '10, pages 9–14. ACM, 2010. 13, 14

[42] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human Mobility, Social Ties, and Link Prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. ACM, 2011. 11

[43] Y. Zhang and M. Pennacchiotti. Predicting Purchase Behaviors from Social Media. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1521–1532, 2013. 1, 10