



# Predicting trading interactions in an online marketplace through location-based and online social networks

Lukas Eberhard<sup>1</sup> · Christoph Trattner<sup>2</sup>  · Martin Atzmueller<sup>3</sup>

Received: 11 November 2017 / Accepted: 27 June 2018 / Published online: 9 July 2018  
© Springer Nature B.V. 2018

## Abstract

Link prediction is a prominent research direction e.g., for inferring upcoming interactions to be used in recommender systems. Although this problem of predicting links between users has been extensively studied in the past, research investigating this issue simultaneously in multiplex networks is rather rare so far. This is the focus of this paper. We investigate the extent to which trading interactions between sellers and buyers within an online marketplace platform can be predicted based on three different but overlapping networks—an online social network, a location-based social network and a trading network. In particular, we conducted the study in the context of the virtual world Second Life. For that, we crawled according data of the online social network, user information of the location-based social network obtained by specialized bots, and we extracted purchases of the trading network. Overall, we generated and used 57 topological and homophilic features in different constellations to predict trading interactions between user pairs. We focused on both unsupervised as well as supervised learning methods. For supervised learning, we achieved accuracy values up to 92.5%, for unsupervised learning we obtained nDCG values up to over 97% and MAP values up to 75%.

**Keywords** Seller · Buyer · Link prediction · Location-based and online social networks · Second life · Supervised and unsupervised learning

---

✉ Christoph Trattner  
trattner.christoph@gmail.com

Lukas Eberhard  
lukas.eberhard@tugraz.at

Martin Atzmueller  
m.atzmuller@uvt.nl

<sup>1</sup> Institute of Interactive Systems and Data Science, TU Graz, Graz, Austria

<sup>2</sup> Department of Information Science and Media Studies, University of Bergen, Bergen, Norway

<sup>3</sup> Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, The Netherlands

## 1 Introduction

Social networks capture useful information about the relations between their users and their social characteristics (Coleman 1988). Since also the evolution of that structure is of particular interest, a large part of recent research activity in social networks is related to the link prediction problem: Here the goal is to estimate, whether two users  $u$  and  $v$  will interact with each other in the future or not (Liben-Nowell and Kleinberg 2007). Most of the work in this area being applied e.g., for friend recommendation (Barbieri et al. 2014), or community recommendation (Backstrom and Leskovec 2011). Recent approaches also include multiple relations that cover social networks from different perspectives, e.g., considering different relations between the set of actors, in order to enable context-aware social personalization and recommendation systems (Eirinaki et al. 2018). However, work that provides insights in which source of information is the most useful one, what types of features shall be used, and how well do both of these perform in unsupervised as well as supervised settings, are rare.

*Objective* The problem addressed in this paper is a particular kind of link prediction problem—relevant both for virtual as well as physical social networks. Here, we want to show if and 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 three sources of data: (1) an online social network, (2) a location-based social network and (3) a trading network, including topological and homophilic features of these three different networks. Here, we aim at extending the existing analyses in two directions: we focus on a combined feature-based analysis of the different networks, in contrast to, for example, Guo et al. (2011) or Zhang and Pennacchiotti (2013), we specifically focus on the impact of different features since they provide actionable insights that can be used for decision making later. We aim at predicting trading interactions between users from four perspectives: an online social, a location-based social, a trading network and different combinations of them in order to see, if this increases our prediction results. We also consider both unsupervised as well as supervised methods, applying different constellations of features generated by the set of networks. This is useful for estimating performance indicators towards real application (Bischoff 2012), and it also provides additional support with respect to the importance of features (and their combinations).

As a data source for our experiments we rely on Second Life<sup>1</sup>: it aims to provide users a platform similar to the real-world but virtually to interact with each other via a social network called My Second Life. This also allows to create businesses over the Second Life marketplace that is similar to what we refer to as ebay<sup>2</sup> in the real world, as shown in Szell et al. (2012), Lehdonvirta (2009) and Guo et al. (2011).

*Research Questions* To drive our research we have defined the following three high-level research questions, which we will investigate and discuss in the following sections:

*RQ1* First, we focus on individual features: to what extent can trading interactions be predicted based on features from a set of networks (social, location-based, trading) individually, considering local (proximity), path-based and content-based features on the (overall) prediction accuracy?

---

<sup>1</sup> <http://secondlife.com>.

<sup>2</sup> <http://ebay.com>.

- RQ2* Second, we focus on the different feature types, as well as their collective interplay: does the combination of different feature types across multiple networks (social, location-based, trading) increase the results of predicting trading interactions?
- RQ3* Third, we focus on different classification approaches and address the question about on different prediction approaches—unsupervised as well as supervised learning methods: what is their impact with respect to different constellations of the available features?

*Outline* In the following sections we will review appropriate background literature, introduce the datasets and methodology chosen to address our research questions, and present and discuss the results of our study. Finally, we draw conclusions, discuss the limitations of our study and propose future research directions.

## 2 Background

Link prediction is a prominent method for link analysis in social networks. It aims at predicting *new* and *recurring* links between the involved actors (Liben-Nowell and Kleinberg 2003, 2007; Getoor and Diehl 2005; Al Hasan and Zaki 2011; Zhang and Philip 2014; Wang et al. 2015; Martínez et al. 2016). However, little work has been done in the context of predicting interactions between sellers and buyers.

Below, we first discuss related approaches for link prediction outlining both unsupervised as well as supervised approaches, before we describe the relation to feature engineering for link prediction: this includes network proximity as well as path-based and content-based measures. According to these dimensions, we summarize the different foci of link prediction approaches in Table 1. The table shows the underlying (learning) method for link prediction, as well as the respective applied features—according to the categories outlined above. We will discuss these in more detail in Sect. 2.2.

This section concludes with a final subsection, where we outline current gaps in the literature, and summarize differences to previous research. Furthermore, we provide and discuss a detailed list of contributions of this work.

### 2.1 Link prediction methods

The prediction of (new) links between nodes in a social network is an interesting and challenging task. A first comprehensive fundamental analysis was done by Liben-Nowell and Kleinberg (2003). In particular, Liben-Nowell and Kleinberg (2003, 2007) 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, using network proximity measures. Such link predictions could be used for suggesting promising interactions between two individuals in such a social network (Zhang et al. 2013; Barbieri et al. 2014). This work is concerned with the prediction of trading interactions using several user information sources similar to Guo et al. (2011).

In the link prediction literature, typically two different types of learning approaches are distinguished: unsupervised and supervised approaches. In the following two paragraphs, we review literature in these two strands of research.

*Unsupervised approaches* Extending the fundamental work of Liben-Nowell and Kleinberg (2003, 2007) using network proximity and path-based measures, Murata and Moriyasu

**Table 1** Overview on (general) unsupervised and supervised methods for link prediction, as well as the utilized sets of features, according to the categorization in feature engineering discussed below

Author(s)	Learning type	Method	Topological	Path based	Homophilic
Backstrom and Leskovec (2011)	Supervised	DT, LR, SRW	×	×	
Barbieri et al. (2014)	Supervised	JSVD, WTFW			×
Cheng et al. (2011)	Supervised	DT, LR	×	×	
Cranshaw et al. (2010)	Supervised	AB, RF, SVM	×		
Eberhard and Trattner (2016)	Unsupervised	CF	×	×	
Fire et al. (2011)	Supervised	AB, ANN, B, C4.5, kNN NB, SVM, RF, RoF	×		
Fire et al. (2013)	Supervised	B, J48, RF	×	×	
Hasan et al. (2006)	Supervised	B, DT, kNN, MP, NB, RBFN, SVM	×		
Jones et al. (2013)	Supervised	RF, SVM	×		×
Kibanov et al. (2015)	Unsupervised	CB	×		×
Leskovec et al. (2010)	Supervised	LR	×		
Liben-Nowell and Kleinberg (2007)	Unsupervised	CF	×	×	
Lichtenwaller et al. (2010)	Un-/supervised	CF/B, J48, NB	×	×	
Lichtenwaller and Chawla (2011)	Un-/supervised	CF/any in WEKA <sup>a</sup>	×	×	
Lichtenwaller and Chawla (2012)	Supervised	B	×		
Lu et al. (2010)	Un-/supervised	Hybrid	×	×	
Lü and Zhou (2010)	Unsupervised	CF	×		
Menon and Eikan (2011)	Supervised	MF	×	×	
Murata and Moriyasu (2007)	Unsupervised	CF	×		
Rowe et al. (2012)	Supervised	LR	×		×
Scellato et al. (2011)	Supervised	J48, MT, NB, RF	×		
Scholz et al. (2013a)	Unsupervised	URW	×		×
Scholz et al. (2012)	Unsupervised	CF	×	×	
Scholz et al. (2013b)	Unsupervised	CF, URW	×	×	
Steurer and Trattner (2013a)	Un-/supervised	CF/J48, LR, SVM	×		×
Steurer and Trattner (2013c)	Supervised	LR, RF, SVM	×		×

**Table 1** (continued)

Author(s)	Learning type	Method	Topological	Path based	Homophilic
Thiele et al. (2018)	Supervised	SM	X		
Zhang and Penacchiotti (2013)	Supervised	LR, NB, SVM			
Zhang et al. (2013)	Supervised	SVM	X	X	X
Zhuang et al. (2012b)	Semi-supervised	PLP-FGM, SVM, TPFPG	X		

<sup>a</sup> Any supervised method available in WEKA, AB AdaBoost, ANN artificial neural networks, B bagging, CB content-based filtering, CF collaborative filtering, DT decision tree, J48 Pruned or Unpruned C4.5 Decision Tree, JSVD joint singular value decomposition, kNN k-nearest-neighbors, LR logistic regression, NB naive bayes, MF matrix factorization, MP multilayer perceptron, MT model trees, PLP-FGM partially-labeled pairwise factor graph model, RBFN radial basis function network, RF random forest, RoF rotation forest, SRW supervised random walk, SVM support vector machine, SM sienna model, TPFPG time-constrained probabilistic factor graph model, URW unsupervised random walk, WTFW Who to follow and Why

(2007) investigated weighted variants of the network proximity measures Adamic–Adar, common neighbors and preferential attachment; essentially these methods obtain a ranking utilizing collaborative filtering techniques for obtaining similar users for a given user. Furthermore, Lü and Zhou (2010) presented an approach to analyze the role of weak ties in social networks, while Zhuang et al. (2012b) used active learning for inferring social ties.

Most of these works analyzed the predictability of new links in online social networks like co-authorship in DBLP or arXiv.org. The prediction of new links in real-world social contacts has been largely neglected. Zhuang et al. (2012a) present prediction techniques using location-based proximity as a weak-proxy for face-to-face encounters and online social networks. In contrast, Scholz et al. (2012) conducted a first analysis concerning the predictability of new links in real face-to-face contact networks. In Scholz et al. (2013a), a method for link prediction on multiplex networks, based on the idea of link prediction using the rooted PageRank algorithm (Liben-Nowell and Kleinberg 2007) is described yielding the Hybrid Rooted PageRank algorithm. This algorithm enables a combined inference on the multiplex network for prediction. In a similar setting, Kibanov et al. (2015) apply content-based filtering.

Also, a structural view on link prediction is taken in Scholz et al. (2013b). In addition, the integration of heterogeneous information for link prediction is investigated in Scholz et al. (2014). Furthermore, Lichtenwalter et al. (2010) as well as Lichtenwalter and Chawla (2011) introduce a novel unsupervised method, i.e., a restricted variant of rooted PageRank, and a new supervised method (Lichtenwalter and Chawla 2012) for link prediction. Here, we extend these approaches covering both supervised and unsupervised methods.

*Supervised approaches* In the literature supervised learning is commonly used to predict links between users in a network whenever label information is available, employing various machine learning methods.

Hasan et al. (2006) considered a social network with interactions as edges representing the co-authoring of research articles. Each article included at least author information and publication year. For 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 training 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. Here, they mainly focused on topological (proximity) features.

Backstrom and Leskovec (2011) introduced a supervised method, based on supervised random walks for predicting new links, focusing on the network structure as well. Similarly, Menon and Elkan (2011) present a supervised approach using matrix factorization. Lu et al. (2010) applied a supervised approach using multiple sources, focusing on feature engineering methods. Compared to these approaches, we provide a much more feature rich setting integrating multiple relations and feature sets into our prediction approach. In addition, Scellato et al. (2011) present a supervised learning framework integrating place features on location-based social networks.

In this paper, we extend the approaches mentioned above by not only focusing on network (proximity) features or integrating place features, but by taking a more comprehensive view: We focus on user, homophilic and locational features, integrating them in order to assess their impact and efficacy. Thiele et al. (2018) present a longitudinal analysis of social network data for link prediction in the scope of the predictive impact of initial

face-to-face contacts on the formation and evolution of developmental peer network relationships. For the predictive model, a longitudinal RSiena model is applied, cf. Ripley et al. (2011). In the context of this paper, we instead focus on the relations between the different networks, considering the combinations of features for predicting buyer–seller interactions.

Overall, several machine learning algorithms for the supervised link prediction in online social networks have been investigated in the past. For example, decision trees [C4.5 (Quinlan 1986), in the J48 implementation of Weka (Hall et al. 2009)] were used in Cheng et al. (2011) and Fire et al. (2011), logistic regression (LR) in Cheng et al. (2011), Jones et al. (2013), Leskovec et al. (2010), Rowe et al. (2012) and Zhang and Pennacchiotti (2013), random forest (RF) in Fire et al. (2013) and Jones et al. (2013), and support vector machines (SVM) in Fire et al. (2011), Hasan et al. (2006), Jones et al. (2013) and Zhang and Pennacchiotti (2013), for a variety of tasks, e.g., reciprocal links, links of new users, or follower connections.

## 2.2 Features for predicting links between users in networks

Within social networks, important information about users and their relations can be extracted in order to assess similarities between users. Topological and homophilic features are hypernyms for such user similarities in partly large-scale network data (Coleman 1988; Steurer and Trattner 2013a). Then, these can be leveraged in link prediction approaches, relying on the (similar) social context of the users, e.g., Liben-Nowell and Kleinberg (2003) and Guo et al. (2011).

*Topological features* If the structure of a network is known, then network topological features can be applied for estimating the similarity between two users in the network, also in longitudinal analysis, e.g., Thiele et al. (2018).

For the analysis of co-authorship social networks, Liben-Nowell and Kleinberg (2007) used topological features for link prediction. They used measures such as common neighbors (number of neighbors that two users have in common), Jaccard's coefficient (number of common divided by number of total neighbors) as proposed by Salton and McGill (1983), Adamic–Adar (regarding the node degree of the common neighbors) proposed by Adamic and Adar (2003) or preferential attachment (multiplication of numbers of neighbors of two users) proposed by Barabasi and Albert (1999), Newman (2001) and Barabasi et al. (2002). More detailed topological feature measures were used by Steurer and Trattner (2013a). They partly used a directed network for their experiments and thus distinguished between outgoing and incoming network topological features; e.g., 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, Fire et al. (2013) defined topological features such as transitive friends (number of outgoing neighbors of a user intersected by the number of incoming neighbors of another user), Katz measure (path oriented measure) proposed by Katz (1953a), opposite direction friends (reciprocity between two users) for directed graphs or shortest paths.

*Path-based features* As an extension of topological features that focus on the (local) neighborhood of a node, path-based features take into account richer connectivity information. The *rooted PageRank* (Liben-Nowell and Kleinberg 2003) algorithm, as an adaption of the PageRank algorithm (Brin and Page 1998) provides the stationary probability distribution sets of nodes regarding a specific starting node, providing a ranking for link prediction. Similarly, the Katz (1953b) measure also takes into account longer paths extending

the neighborhood, weighted by a damping factor. Katz basically measures the strength of the connection between two nodes: The more paths two nodes are connected with and the shorter these paths are, the stronger the connection. For very small values of those, Katz is actually similar to a network proximity measure based on the nodes' neighborhood, because path lengths greater than three do not contribute very much (Liben-Nowell and Kleinberg 2003).

*Homophilic features* Thelwall (2009) 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 relationships. 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, as investigated by McPherson et al. (2001) in the context of online social networks. Steurer and Trattner (2013a) used attributes as groups, interests, user interactions, events and regions for the computation of homophilic features for their experiments for predicting partnerships in social networks. For the different attributes they computed measures—also used in this paper—such as common items (number of items of an attribute two users have in common), total items (number of total items of an attribute of two users), Jaccard's coefficient (common items divided by the number of total items) or cosine similarity of the item vectors.

*Feature set modeling* Overall, we can approach the link prediction task using different feature sets, dependent on their availability, both individually as well as in combination covering multiple networks. Although Fire et al. (2011) 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. In order to obtain topological network features, the previous 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, e.g., as shown by Thelwall (2009). They attained highly significant indications of homophily for measures such as ethnicity, age, religion, sexual orientation, country or marital status for their experiments with a MySpace dataset. Furthermore, regarding the work of Cranshaw et al. (2010), they applied a hybrid approach, combining 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 (2013a) also combined online social network data with location-based social network data in their partnership prediction experiments.

### 2.3 Differences to previous research and contributions

In summary, the background literature discussed above clearly shows that the general link prediction problem is a well-studied area of research. Many studies have been performed to predict links in online social networks or other types of networks. However,



surprisingly little work has been conducted employing several different networks at the same time, e.g., concerning different types of networks, and in particular the combination of multiple networks, as well as feature types.

We started to investigate that topic using a location-based and an online social network in Steurer et al. (2013), Steurer and Trattner (2013c) and Kibanov et al. (2015) regarding interactions and their types. In addition, we tackled link prediction in the context of multiplex networks for predicting face-to-face interactions in Scholz et al. (2013a), and for attending talks in the context of academic conferences Scholz et al. (2014).

Furthermore, when reviewing the literature regarding the particular problem we study, namely predicting links (trading interactions) between sellers and buyers, we see that actually very little work can be found for that particular area. To the best of our knowledge there is only one study that is directly comparable to our work that has been performed by Guo et al. (2011) in the past, apart from preliminary own work of authors of this paper (Eberhard and Trattner 2016).

The work of Guo et al. (2011) is interesting as it is the first to study usefulness of social networks and 13 different features to predict seller buyer interactions. The context of their work is the largest electronic marketplace in China named TAOBAO, with over 370 million registered users at the end of 2010. Among the features investigated, they employ centrality metrics, such as PageRank as well as homophilic metrics such as the number of common friends a seller and a buyer have in common to predict trading interactions. Also they used the prizes of the products as well as the ratings of the products as a proxy. The intention behind this is, that buyers typically buy from popular sellers (captured, for example, over centrality metrics such as PageRank) or keep also others types of types of relations, such as common friends that have been buying an item from the buyer before.

In this work we use similar features. However, compared to the work of Guo et al. (2011) we do not only rely on social network data, but also reveal whether there is also a signal present in the people's location-based network to predict trading interactions. The features engineered in our approach are based on the existing related work. This includes link-prediction and recommender systems research as well as sociology as mentioned before, but also economy (DiMaggio and Louch 1998) which suggests that the social embeddedness of the sellers in the buyers networks is inherently important for further purchase decisions. As such, we induce in total 57 different features capturing not only network effects between sellers and buyers, but we also consider homophilic features such as the “number of groups or interests” that they have in common.

Furthermore, we make use of location-based network data, to understand whether features such as, for example, the “number of times seller and buyers have been seen in the same location” bears a signal that can be exploited to predict seller-buyer links in the future. In addition to this, our experiments make use of different supervised and unsupervised learning approaches.

To the best of our knowledge, this is the first kind of study that shows the relation between and individual as well as combined impact of three different types of networks, features, features sets and learning methods to predict links between sellers and buyers.

In total, the contributions of this work can be summarized as follows:

1. The collection of a unique dataset of user and activity data in three different networks: an online social network, a location-based social network and a trading network.

2. The engineering of 57 topological and homophilic features to predict trading interactions between two users in these three different types of networks.
3. The statistical analysis of differences of the features to discriminate between seller and buyer interactions across different networks.
4. The presentation of a set of supervised and unsupervised learning experiments to show the meaningfulness of the inducted features individually to predict trading interactions between sellers and buyers.
5. The presentation of results revealing the value of predicting seller buyer trading interactions based on feature sets (homophilic and topological) as well as different kinds of networks (trading, social and location-based).
6. Finally, we show how the features correlate with each other and reveal their importance (measured via Information Gain) when considering all features at the same time in the model.

### 3 Datasets

In order to address the three research questions, it was necessary to have three different kinds of data available:

1. Data from an online social network such as Facebook<sup>3</sup> or Google+<sup>4</sup> where users share personal information on their profiles and are able to communicate with others via the platform.
2. Data from a location-based social network such as Foursquare,<sup>5</sup> where geographical position information about the users is available.
3. Data from a trading network such as eBay,<sup>6</sup> where the community is able to trade with goods. The majority of these websites 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.

As a consequence, we opted for the virtual world Second Life, which unites all the required kinds of data sources. On top of that, there is the advantage of a large user overlap across the three parts of the Second Life platform (Steurer and Trattner 2013a, b, c; Steurer et al. 2013). Although the users in Second Life do not interact with their real life names, but with the names of their avatars in a virtual world, La and Michiardi (2008) and Varvello et al. (2008) have shown that the avatars' behavior tends to be similar to the behavior of humans.

The basic principle of Second Life is that avatars explore the virtual world, meet other avatars and communicate, play or trade with them. Varvello and Voelker (2010) denoted the Second Life social network as small-world network and much more similar to a real world network in comparison with popular online social networks. Crucial for this observation is the establishing of social relationships between users in Second Life, which requires

---

<sup>3</sup> <http://facebook.com/>.

<sup>4</sup> <http://plus.google.com/>.

<sup>5</sup> <http://foursquare.com/>.

<sup>6</sup> <http://ebay.com/>.

**Table 2** Extracted entities from the three different data sources

Online social network	Location-based social network	Trading network
Users	Users	Users
Interactions	Events	Product categories
Postings	Event categories	Product prices
Text messages	Event regions	Product ratings
Pictures		
Comments		
Loves		
Groups		
Interests		
Check-ins		
Favored regions		

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 such as text messages between the users.

As source for our online social network we crawled data from My Second Life.<sup>7</sup> The location-based data were monitored in-world in Second Life and the Second Life Marketplace<sup>8</sup> was used as trading network in this paper. Table 2 provides an overview of the extracted entities of the respective data source that we further used for the feature engineering.

### 3.1 Online social network data

As described in our previous work (Eberhard and Trattner 2016), users in the virtual world of Second Life are able to establish social links through an online social networking platform called My Second Life. Similar to Facebook and its timeline, My Second Life gives Second Life users the opportunity to present personal information on their user profiles or to interact with other users on the so-called feed. Apart from such information about the Second Life avatar such as interests, the day of birth in Second Life, or the biography, users are able to join groups or to 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. A “love” in Second Life is similar to a “like” in Facebook or a plus in Google+. A considerable difference to Facebook exists concerning friendship relations. Such a relation type does not exist in My Second life (Steurer and Trattner 2013a).

Based on the crawling methodology described in our previous work (Trattner and Steurer 2015), at the end of March 2013 we crawled the Second Life profiles of users with public user profiles. We extracted a list of user names from the location-based dataset (see Sect. 3.2) and iteratively extended it by further users who interacted on the feed with the users from the list. For each user, we obtained their interests, the joined groups and the feed

<sup>7</sup> <http://my.secondlife.com/>.

<sup>8</sup> <http://marketplace.secondlife.com/>.

**Table 3** Basic statistics of the online social network dataset

Online social network $G_O$	
Num. users	152,509
Num. edges	270,567
Type	Directed
Degree	3.55
Num. connected components	13,115
Largest connected component	77.69%
Num. postings (text messages / pictures)	226,668
Num. comments	348,106
Num. loves	1,494,044
Num. overall interactions	2,068,818
Average num. interactions per user	$\approx 14$
Num. group joins	1,869,281
Num. unique groups	204,769
Num. users with group join(s)	114,205
Num. stated interests	227,596
Num. unique interests	62,170
Num. users who stated interest(s)	36,610
Num. check-ins	466,930
Num. unique checked-in regions	13,251
Num. users with check-ins	36,430
Num. stated favored regions	337,732
Num. unique favored regions	22,742
Num. users who stated favored region(s)	76,093

interactions with others. Two different sources of Second Life regions were also part of the collected information for each user. In Second Life it is possible to record in-world snapshots of regions in terms of pictures and share them on the feed to show others where users have actually been at a particular time. We collected these so-called check-ins 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.

We constructed the online social network on the basis of the feed interactions between the users, as an indicator for being acquainted. If the number of interactions was zero, no link was generated between them. Users with numbers of interactions greater than or equal to 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 feeds. 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. Due to the fact that this paper 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 of the online social network

**Table 4** Basic statistics of the location-based social network dataset

Location-based social network $G_L$	
Num. users	122,936
Num. edges	1,414,389
Type	Undirected
Degree	23.01
Num. connected components	719
Largest connected component	98.01%
Num. events entries	1,966,206
Num. unique events	81,671
Num. event categories	11
Average num. events per user	$\approx 16$
Num. event regions entries	16,375,540
Num. unique event regions	3,972
Average num. entries per user	$\approx 133$

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 to other users' postings and not one's own. Furthermore, the average of the number of interests defined by each user was  $\approx 1.5$  and the joined groups per user  $\approx 12.4$  on average. On average, for each user we obtained  $\approx 3.1$  check-ins on the feed and  $\approx 2.2$  favored regions. Table 3 gives an overview of the numbers of the online social dataset.

### 3.2 Location-based social network data

We extracted the location-based dataset used in this paper from the in-world of Second Life by scripted robots collecting information about surrounding users. As described in our previous work (Trattner and Steurer 2015), we sent the bots to locations of the Second Life event calendar from the Second Life website to presumably target regions with a higher user frequency than in other places in the huge world of Second Life. Starting from March 2012, over the period of one year, the collected user information formed the basis for the location-based social network. Overall, nearly 19 million data entries with 410,619 different users in 4146 different locations were observed. To generate a network with an adequate density from this huge amount of data, we created a link between two users, if they had met each other more than only once. This is formally defined as  $G_L = \langle V_L, E_L \rangle$ , where  $V_L$  is 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. This rule reduced the number of edges in this network many times over to 1,414,389 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 11 different event categories—an average of  $\approx 16$  events per user. There were 16,375,540 event regions entries registered with 3972 unique regions,

**Table 5** Basic statistics of the trading network dataset

Trading network $G_T$	
Num. users	87,300
Num. edges	219,889
Type	Directed
Degree	5.04
Num. connected components	933
Largest connected component	97.39%
Num. sellers	17,914
Num. buyers	77,645
Num. sellers/buyers	8259
Num. product categories	22
Num. products	120,762
Average num. products per seller	$\approx 7$
Num. purchases	268,852
Average num. purchases per seller	$\approx 15$
Average num. purchases per buyer	$\approx 3$

which means that on the average, each user was found  $\approx 133$  times by the bots. Table 4 provides an overview on the location-based dataset.

### 3.3 Trading network data

Besides the in-world of Second Life and My Second Life, there is an online trading platform called Second Life Marketplace where Second Life users are able to trade with virtual goods. The users can act as sellers, buyers, or both, similarly to common online shopping platforms such as eBay. 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 our basic idea for the trading network for the experiments in this paper.

Based on the crawling methodology described in our previous work (Eberhard and Trattner 2016), we gathered all store sites of the Second Life Marketplace with a web crawler to collect the purchase information. This 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, 8259 users acted as both seller and buyer. The total number of involved users was 87,300. An overview of the trading network dataset is provided in Table 5.

**Table 6** Basic statistics of the combined network

Combined network $G_C$	
Num. users	10,420
Num. online social network edges	8543
Num. location-based social network edges	45,558
Num. trading network edges	5376
Total num. edges	59,477
Num. sellers	2086
Num. buyers	9655
Num. sellers/buyers	1321

## 4 Methodology

In the previous sections we introduced the different sources of data used in our experiments. In this section we describe the methodology for the experiments used in this paper to answer the research questions.

### 4.1 Dataset pre-processing

To make the results for the different networks comparable, it was necessary to bring them on a common basis. We intersected the online social network, the location-based social network, and the trading network by picking out the common nodes of all networks. This means that we considered only those users 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 an in-world observation by the robots.

We formally defined this combined network 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:  $E_C = \{(u, v) \mid (u, v) \in E_O \text{ or } (u, v) \in E_L \text{ or } (u, v) \in E_T, \text{ and } u, v \in V_C\}$ .

The numbers of this combined network, which we used for all the experiments in this paper, are shown in Table 6.

All basic computations for preparing the experiments were done with Python and networkx.<sup>9</sup> The experiments were completed in a way where 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.

### 4.2 Feature engineering

As mentioned in the background section (see Sect. 2), different characteristics can be extracted from networks and used for predicting links between users. In the following

<sup>9</sup> <http://networkx.github.io/>.

three subsection we describe in detail the features and feature sets which have been used and engineered to predict seller-buyer trading interactions. The first subsection describes how we inducted features from the online social network features, followed by a subsection detailing on how we induced features from the location-based social network. Finally, the trading network features are introduced. For each of the three data sources, we derive topological and homophilic features. While the former means inducing network-specific features the latter refers to features extracted from content.

#### 4.2.1 Online social network features

*Topological features* We defined the neighbors of a user  $u$  in this directed network 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 (Steurer and Trattner 2013a). We denoted the definition of outgoing neighbors of a user  $u \in V_O$  as  $\Theta^+(u) = \{v \mid (u, v) \in E_O\}$  and incoming neighbors as  $\Theta^-(u) = \{v \mid (v, u) \in E_O\}$ . Therefore, we could formally compute the whole set of neighbors of  $u$  as  $\Theta(u) = \Theta^+(u) \cup \Theta^-(u)$ .

- *Num. common outgoing neighbors* We defined the number of neighbors that two users  $u$  and  $v$  have in common related to the outgoing communication of them 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  $w$ .
- *Num. 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. We defined this feature 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  $v$ .
- *Outgoing Jaccard's coefficient* The Jaccard's coefficient is the division of the number of common by the number of total neighbors of two users  $u$  and  $v$  and could be seen as a measure for exclusiveness of the relation between them (Cranshaw et al. 2010). It was also split into an outgoing and an incoming feature. We denoted the outgoing Jaccard's coefficient 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 we defined it as the number of common incoming neighbors divided by the number of total incoming neighbors:  $O_{JC}^-(u, v) = \frac{|\Theta^-(u) \cap \Theta^-(v)|}{|\Theta^-(u) \cup \Theta^-(v)|}$ .
- *Preferential attachment+–* Here the preferential attachment score, first proposed by Barabasi and Albert (1999), is presented in a slightly different way, proposed by Cheng et al. (2011). It is another popular measure to describe the correlation between the out-degree of a user  $u$  and the in-degree of a user  $v$ . We calculated the value for this feature 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–+* The difference to the preferential attachment +– feature described above is, that the in- and out-degree of the involved users were swapped. So we denoted the preferential attachment in feature for two users  $u$  and  $v$  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 bidirec-



tional or in only one direction (Cheng et al. 2011). We denoted this feature 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* Regarding the relation between two users related to their neighbors, Adamic and Adar (2003) proposed 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. (2011) suggested a refinement of the Adamic–Adar measure in which only the common incoming neighbors are considered:  $O_{AA}(u, v) = \sum_{z \in \theta^-(u) \cap \theta^-(v)} \frac{1}{\log(|\theta(z)^-|)}$ .
- *Katz* Katz is a path-based attribute proposed by Katz (1953a) that measures the strength of the connection between two nodes in a network. The more paths two nodes are connected with and the shorter these paths are, the stronger is the connection between the nodes.  $\beta$  expresses the emphasis of the path length  $l$  between two nodes  $u$  and  $v$ . The weight of shorter path lengths rises by decreasing  $\beta$ . Due to the high complexity calculating this measure for large networks, we introduced a cutoff  $c = 3$  considering only paths with a maximum length of 3. With  $|path_{u,v}^l|$  as the number of paths between  $u$  and  $v$  of length  $l$ , we formally defined the Katz measure as  $O_{K\beta}(u, v) = \sum_{l=1}^c \beta^l \cdot |path_{u,v}^l|$ .
- *Rooted PageRank* The rooted PageRank (Liben-Nowell and Kleinberg 2007) is also a path-based measure and special kind of the personalized PageRank (Chakrabarti 2007). The rooted PageRank score between two nodes  $u$  and  $v$  is based on a random walk on the network starting at node  $u$ . With probability  $1 - \alpha$  it iteratively moves to a random neighbor of the current node and with probability  $\alpha$  it jumps back to  $u$ . We defined it as  $O_{RPR\alpha}(u, v) =$  stationary probability of  $v$  under the following random walk (Pearson 1905; Spitzer 2013): (1) with probability  $1 - \alpha$  move to a random neighbor of the current node, and (2) with probability  $\alpha$  return to  $u$ .

*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  $\Phi(u)$ .

- *Num. common groups* This feature represents the number of groups two users  $u$  and  $v$  have in common:  $G_C(u, v) = |\Delta(u) \cap \Delta(v)|$ .
- *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 we denoted the Jaccard’s coefficient for groups as  $G_{JC}(u, v) = \frac{|\Delta(u) \cap \Delta(v)|}{|\Delta(u) \cup \Delta(v)|}$ .
- *Num. 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:  $I_C(u, v) = |\Phi(u) \cap \Phi(v)|$ .
- *Jaccard’s coefficient of interests* For the user-defined interests we computed the Jaccard’s coefficient for the proportion of common and total interests of two users  $u$  and  $v$  as  $I_{JC}(u, v) = \frac{|\Phi(u) \cap \Phi(v)|}{|\Phi(u) \cup \Phi(v)|}$ .
- *Num. 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. We defined the interactions from a user  $u$  to a user  $v$  as  $\iota(u, v)$ . So this feature shows the number of all interactions from  $u$  to  $v$  and we formally defined it 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  $A(u)$ .

- *Num. common check-ins* This feature is a measure for how many common regions two users  $u$  and  $v$  checked in and shared on their own feed and we formally specified it as  $RR_C(u, v) = |A(u) \cap A(v)|$ .
- *Jaccard's coefficient of check-ins* The value of the common check-ins divided by the value of the total check-ins of two users  $u$  and  $v$  is the Jaccard's coefficient measure again and we defined it as  $RR_{JC}(u, v) = \frac{|A(u) \cap A(v)|}{|A(u) \cup A(v)|}$ .
- *Overlap of check-ins* The overlap of the sets of check-ins 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. We stated this feature as  $RR_O(u, v) = \frac{|A(u) \cap A(v)|}{|A(u)| + |A(v)|}$ .

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 check-ins. We defined the favored regions of a user  $u$  as  $\Xi(u)$ .

- *Num. common favored regions* We defined the number of favored regions two users  $u$  and  $v$  have in common as  $RF_C(u, v) = |\Xi(u) \cap \Xi(v)|$ .
- *Jaccard's coefficient of favored regions* We state the Jaccard's coefficient of favored regions of two users  $u$  and  $v$  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  $RF_O(u, v) = \frac{|\Xi(u) \cap \Xi(v)|}{|\Xi(u)| + |\Xi(v)|}$ .

#### 4.2.2 Location-based social network features

*Topological features* In the location-based social network we defined the neighbors of a user  $u \in V_L$  as  $\Gamma(u) = \{v \mid (u, v) \in E_L\}$ . Similar to the topological online social network features described in Sect. 4.2.1, we subdivided features to measure the structural overlap of two users in the location-based social network as follows:

- *Num. common neighbors* This feature represents the number of neighbors two users  $u$  and  $v$  have in common. We denoted the common neighbors as  $L_{CN}(u, v) = |\Gamma(u) \cap \Gamma(v)|$ .
- *Jaccard's coefficient* We stated the Jaccard's coefficient for two users  $u$  and  $v$  in the location-based social network 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 Sect. 4.2.1, we formally defined the Adamic–Adar for undirected networks as

$$L_{AA}(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log(|\Gamma(z)|)}.$$

- *Num. days seen* We defined the set of days two users were in the same region at the same time as  $\eta$ . Therefore, the number of days two users  $u$  and  $v$  have met each other in-world was formally stated as  $L_{DS}(u, v) = |\eta(u, v)|$ .
- *Mean distance*  $\omega$  is the set of distances two users were apart in the same region at a certain point in time gathered by our bots. We stated this feature as  $L_{MD}(u, v) = \frac{1}{|\omega(u, v)|} \sum_{d \in \omega(u, v)} d$ .
- *Katz* As described for the online social network, Katz is a path-based measure that quantifies the strength of the connection between two nodes in a network based on lengths of the paths between them. For the location-based social network, again with a cutoff  $c = 3$ , we defined it as  $L_{K\beta}(u, v) = \sum_{l=1}^c \beta^l \cdot |\text{path}_{u,v}^l|$ .
- *Rooted PageRank* As mentioned above, the rooted PageRank is the stationary probability of node  $v$  based on a random walk that starts at node  $u$ . It iteratively moves to a random neighbor of the current node or jumps back to  $u$ . We formally denote it as  $L_{RPR\alpha}(u, v)$ , the stationary probability of  $v$  under the following random walk: (1) with probability  $1 - \alpha$  move to a random neighbor of the current node, (2) with probability  $\alpha$  return to  $u$ .

*Homophilic features* As mentioned in Sect. 3.2 the implemented robots monitored users in-world at Second Life events. We stated the events a user  $u$  visited as  $\Phi(u)$ . The following features refer to such events and their locations:

- *Num. common events* We defined the number of common events which two users  $u$  and  $v$  visited as  $E_C(u, v) = |\Pi(u) \cap \Pi(v)|$ .
- *Jaccard's coefficient of events* We computed the Jaccard's coefficient measure of the events two users  $u$  and  $v$  visited 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. We computed the cosine similarity of event categories between two users  $u$  and  $v$  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 check-ins and favored regions:

- *Num. common event regions* We stated the number of regions of events two users  $u$  and  $v$  visited in common as  $RE_C(u, v) = |Y(u) \cap Y(v)|$ .
- *Jaccard's coefficient of event regions* This feature measures the Jaccard's coefficient of the event regions of two users  $u$  and  $v$ :  $RE_{JC}(u, v) = \frac{|Y(u) \cap Y(v)|}{|Y(u) \cup Y(v)|}$ .
- *Overlap of event regions* We defined 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  $v$  in this feature as  $RE_O(u, v) = \frac{|Y(u) \cap Y(v)|}{|Y(u)| + |Y(v)|}$ .

### 4.2.3 Trading network features

*Topological features* The topological features to measure the structural overlap of two users in the online social network described in Sect. 4.2.1 could also be applied on the trading network of the Second Life Marketplace. Since this network is directed, we split some of the features into outgoing and incoming features again.

We denoted the outgoing neighbors in the trading network of a user  $u \in V_T$  as  $\Psi^+(u) = \{v \mid (u, v) \in E_T\}$  and incoming neighbors as  $\Psi^-(u) = \{v \mid (v, u) \in E_T\}$ . The formal definition of the combined set of neighbors is then stated as  $\Psi(u) = \Psi^+(u) \cup \Psi^-(u)$ .

- *Num. common outgoing neighbors* We defined the number of outgoing neighbors two users  $u$  and  $v$  have in common as  $T_{CN}^+(u, v) = |\Psi^+(u) \cap \Psi^+(v)|$ .
- *Num. common incoming neighbors* The definition for the number of common incoming neighbors of two users  $u$  and  $v$  was stated as  $T_{CN}^-(u, v) = |\Psi^-(u) \cap \Psi^-(v)|$ .
- *Outgoing Jaccard's coefficient* We denoted the definition of the outgoing Jaccard's coefficient of two users  $u$  and  $v$  of the trading network 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, given as  $T_{JC}^-(u, v) = \frac{|\Psi^-(u) \cap \Psi^-(v)|}{|\Psi^-(u) \cup \Psi^-(v)|}$ .
- *Preferential attachment +-* 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 we defined it as  $T_{PS}^+(u, v) = |\Psi^+(u)| \cdot |\Psi^-(v)|$ .
- *Preferential attachment -+* The difference to the preferential attachment +- feature is the swapping of the users. We formally defined this feature as  $T_{PS}^-(u, v) = |\Psi^-(u)| \cdot |\Psi^+(v)|$ .
- *Reciprocity of trading interactions* 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, we stated this feature 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. We used the cosine similarity measures for the following features:

- *Cosine similarity of product categories* To compute a value for the similarity between the product categories of a user pair  $(u, v)$ , we defined two vectors  $\zeta(u)$  and  $\zeta(v)$ . 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 Sect. 4.2.2, we computed the cosine similarity of product categories between  $u$  and  $v$  as  $P_{CCos}(u, v) = \frac{\zeta(u) \cdot \zeta(v)}{\|\zeta(u)\| \|\zeta(v)\|}$ .

**Table 7** Overview of all features: formal definition, mean values of user pairs with ( $\emptyset Val_{uv}$ ) and without ( $\emptyset Val_{uv}$ ) trading interactions, and sign

Feature	Description	Formal definition	$\emptyset Val_{uv}$	$\emptyset Val_{uv}$	Sign.
<i>Online social network</i>					
<b>Topological</b>					
$O_{CN}^+$	Num. common outgoing neighbors	$O_{CN}^+(u, v) =  \theta^+(u) \cap \theta^+(v) $	1.16e-01	5.00e-03	
$O_{CN}^-$	Num. common incoming neighbors	$O_{CN}^-(u, v) =  \theta^-(u) \cap \theta^-(v) $	1.56e-01	3.96e-03	
$O_{JC}^+$	Outgoing Jaccard's coefficient	$O_{JC}^+(u, v) = \frac{ \theta^+(u) \cap \theta^+(v) }{ \theta^+(u) \cup \theta^+(v) }$	2.13e-03	1.66e-04	
$O_{JC}^-$	Incoming Jaccard's coefficient	$O_{JC}^-(u, v) = \frac{ \theta^-(u) \cap \theta^-(v) }{ \theta^-(u) \cup \theta^-(v) }$	2.39e-03	2.15e-04	
$O_{PS}^+$	Preferential attachment+	$O_{PS}^+(u, v) =  \theta^+(u)  \cdot  \theta^-(v) $	1.05e+02	1.17e+01	***
$O_{PS}^-$	Preferential attachment-	$O_{PS}^-(u, v) =  \theta^-(u)  \cdot  \theta^+(v) $	9.41e+01	1.45e+01	***
$O_R$	Reciprocity of user communication	$O_R(u, v) = \begin{cases} 0 & \text{if } (u, v) \in E_O, (v, u) \notin E_O \\ 1 & \text{if } (u, v) \in E_O, (v, u) \in E_O \end{cases}$	1.86e-02	3.72e-05	*
$O_{AA}$	Adamic-Adar	$O_{AA}(u, v) = \sum_{z \in \theta^-(u) \cap \theta^-(v)} \frac{1}{\log( \theta^-(z) )}$	1.02e-01	1.94e-03	
$O_{K01}$	Katz ( $\beta = .001$ )	$O_{K\beta}(u, v) = \sum_{l=1}^c \beta^l \cdot  path_{l,u,v}^+ $	1.55e-05	4.06e-08	*
$O_{K01}$	Katz ( $\beta = .01$ )		1.72e-04	9.58e-07	*
$O_{K1}$	Katz ( $\beta = .1$ )		9.79e-03	3.08e-04	*
$O_{RPR01}$	Rooted PageRank ( $\alpha = .01$ )	$O_{RPR\alpha}(u, v)$ = stationary probability of v/random walk: (1) with probability $1 - \alpha$ move to a random neighbor of current node, (2) with probability $\alpha$ return to $u$	1.79e-03	1.64e-05	*
$O_{RPR05}$	Rooted PageRank ( $\alpha = .05$ )		1.91e-03	1.92e-05	*
$O_{RPR15}$	Rooted PageRank ( $\alpha = .15$ )		1.98e-03	2.41e-05	*
$O_{RPR3}$	Rooted PageRank ( $\alpha = .3$ )		1.95e-03	3.17e-05	*
$O_{RPR5}$	Rooted PageRank ( $\alpha = .5$ )		1.70e-03	4.25e-05	*
<b>Homophilic</b>					
$G_C$	Num. common groups	$G_C(u, v) =  A(u) \cap A(v) $	2.43e-01	6.06e-02	***

**Table 7** (continued)

Feature	Description	Formal definition	$\emptyset Val_v$	$\emptyset Val_{hp}$	Sign.
$G_{JC}$	Jaccard's coefficient of groups	$G_{JC}(u, v) = \frac{ A(u) \cap A(v) }{ A(u) \cup A(v) }$	6.33e-03	1.25e-03	***
$I_C$	Num. common interests	$I_C(u, v) =  \Phi(u) \cap \Phi(v) $	2.01e-02	1.05e-02	
$I_{JC}$	Jaccard's coefficient of interests	$I_{JC}(u, v) = \frac{ \Phi(u) \cap \Phi(v) }{ \Phi(u) \cup \Phi(v) }$	1.14e-03	6.51e-04	
$OI$	Num. interactions	$OI(u, v) =  I(u, v) $	4.64e-01	9.30e-05	
$RR_C$	Num. common check-ins	$RR_C(u, v) =  A(u) \cap A(v) $	1.12e-02	6.88e-04	
$RR_{JC}$	Jaccard's coefficient of check-ins	$RR_{JC}(u, v) = \frac{ A(u) \cap A(v) }{ A(u) \cup A(v) }$	2.13e-04	5.39e-05	
$RR_O$	Overlap of check-ins	$RR_O(u, v) = \frac{ A(u) \cap A(v) }{ A(u)  +  A(v) }$	1.81e-04	4.11e-05	*
$RF_C$	Num. common favored regions	$RF_C(u, v) =  \Xi(u) \cap \Xi(v) $	3.55e-02	3.35e-03	*
$RF_{JC}$	Jaccard's coefficient of favored regions	$RF_{JC}(u, v) = \frac{ \Xi(u) \cap \Xi(v) }{ \Xi(u) \cup \Xi(v) }$	6.50e-03	4.74e-04	*
$RF_O$	Overlap of favored regions	$RF_O(u, v) = \frac{ \Xi(u) \cap \Xi(v) }{ \Xi(u)  +  \Xi(v) }$	5.08e-03	3.72e-04	*
<i>Location-based social network</i>					
Topological					
$L_{CN}$	Num. common neighbors	$L_{CN}(u, v) =  I(u) \cap I(v) $	2.57e+00	2.38e-01	***
$L_{JC}$	Jaccard's coefficient	$L_{JC}(u, v) = \frac{ I(u) \cap I(v) }{ I(u) \cup I(v) }$	1.02e-02	1.02e-03	***
$L_{AA}$	Adamic-Adar	$L_{AA}(u, v) = \sum_{z \in I(u) \cap I(v)} \frac{1}{\log(I(z))}$	1.58e+00	1.11e-01	***
$L_{DS}$	Num. days seen	$L_{DS}(u, v) =  H(u, v) $	3.14e-01	3.66e-03	**
$L_{MD}$	Mean distance	$L_{MD}(u, v) = \frac{1}{ o(u, v) } \sum_{d \in o(u, v)} d$	4.22e-01	3.54e-02	**
$L_{K001}$	Katz ( $\beta = .001$ )	$L_{K\beta}(u, v) = \sum_{l=1}^c \beta^l \cdot  path^l_{u,v} $	3.80e-05	9.82e-07	***
$L_{K01}$	Katz ( $\beta = .01$ )		8.40e-04	6.79e-05	***
$L_{K1}$	Katz ( $\beta = .1$ )		2.60e-01	3.94e-02	***

**Table 7** (continued)

Feature	Description	Formal definition	$\emptyset Val_v$	$\emptyset Val_{wv}$	Sign.
$L_{RPR01}$	Rooted PageRank ( $\alpha = .01$ )	$L_{RPR\alpha}(u, v) =$ stationary probability of $v$ /random walk: (1) with probability $1 - \alpha$ move to a random neighbor of current node, (2) with probability $\alpha$ return to $u$	4.33e-04	2.82e-05	***
$L_{RPR05}$	Rooted PageRank ( $\alpha = .05$ )		6.99e-04	3.15e-05	***
$L_{RPR15}$	Rooted PageRank ( $\alpha = .15$ )		1.07e-03	3.92e-05	***
$L_{RPR3}$	Rooted PageRank ( $\alpha = .3$ )		1.29e-03	5.07e-05	***
$L_{RPR5}$	Rooted PageRank ( $\alpha = .5$ )		1.23e-03	6.64e-05	***
<b>Homophilic</b>					
$E_C$	Num. common events	$E_C(u, v) =  \Pi(u) \cap \Pi(v) $	3.45e-01	6.92e-03	***
$E_{JC}$	Jaccard's coefficient of events	$E_{JC}(u, v) = \frac{ \Pi(u) \cap \Pi(v) }{ \Pi(u) \cup \Pi(v) }$	7.75e-03	1.78e-04	***
$E_{CCos}$	Cosine similarity of event categories	$E_{CCos}(u, v) = \frac{\delta(u, \delta(v))}{\ \delta(u)\  \ \delta(v)\ }$	5.19e-01	5.02e-01	*
$RE_C$	Num. Common Event Regions	$RE_C(u, v) =  Y(u) \cap Y(v) $	3.07e-01	1.43e-01	***
$RE_{JC}$	Jaccard's coefficient of event regions	$RE_{JC}(u, v) = \frac{ Y(u) \cap Y(v) }{ Y(u) \cup Y(v) }$	2.74Pe-02	1.04e-02	***
$RE_O$	Overlap of event regions	$RE_O(u, v) = \frac{ Y(u) \cap Y(v) }{ Y(u)  +  Y(v) }$	2.19e-02	9.18e-03	***
<b>Trading network</b>					
<b>Topological</b>					
$T_{CN}^+$	Num. common outgoing neighbors	$T_{CN}^+(u, v) =  \Psi^+(u) \cap \Psi^+(v) $	4.99e-02	4.95e-03	*
$T_{CN}^-$	Num. common incoming neighbors	$T_{CN}^-(u, v) =  \Psi^-(u) \cap \Psi^-(v) $	1.00e-01	9.10e-03	***
$T_{JC}^+$	Outgoing Jaccard's coefficient	$T_{JC}^+(u, v) = \frac{ \Psi^+(u) \cap \Psi^+(v) }{ \Psi^+(u) \cup \Psi^+(v) }$	5.82e-04	1.41e-04	***
$T_{JC}^-$	Incoming Jaccard's coefficient	$T_{JC}^-(u, v) = \frac{ \Psi^-(u) \cap \Psi^-(v) }{ \Psi^-(u) \cup \Psi^-(v) }$	3.07e-03	7.05e-04	***
$T_{PS}^+$	Preferential Attachment +	$T_{PS}^+(u, v) =  \Psi^+(u)  \cdot  \Psi^-(v) $	1.69e+03	4.90e+01	***
$T_{PS}^-$	Preferential attachment -	$T_{PS}^-(u, v) =  \Psi^-(u)  \cdot  \Psi^+(v) $	1.51e+01	9.01e+00	***

**Table 7** (continued)

Feature	Description	Formal definition	$\emptyset Val_v$	$\emptyset Val_{hv}$	Sign.
$T_R$	Reciprocity of trading interactions	$T_R(u, v) = \begin{cases} 0 & \text{if } (u, v) \in E_T, (v, u) \notin E_T \\ 1 & \text{if } (u, v) \in E_T, (v, u) \in E_T \end{cases}$	6.32e-03	7.44e-05	
$T_{AA}$	Adamic-Adar	$T_{AA}(u, v) = \sum_{z \in \Psi^{(u)} \cap \Psi^{(v)}} \frac{1}{\log( \Psi^{(z)} )}$	8.23e-02	7.34e-03	**
<b>Homoph.</b>					
$P_{CCos}$	Cosine similarity of product categories	$P_{CCos}(u, v) = \frac{\tau(u) \cdot \tau(v)}{\ \tau(u)\  \ \tau(v)\ }$	4.26e-01	1.95e-01	***
$P_{PCos}$	Cosine similarity of product prices	$P_{PCos}(u, v) = \frac{p(u) \cdot p(v)}{\ p(u)\  \ p(v)\ }$	4.71e-01	2.96e-01	***
$P_{RCos}$	Cosine similarity of product ratings	$P_{RCos}(u, v) = \frac{r(u) \cdot r(v)}{\ r(u)\  \ r(v)\ }$	6.93e-01	5.54e-01	***

(\*\*\* $p < 0.001$ , \*\* $p < 0.01$ , \* $p < 0.1$ ). Edges for the online social network were formed based on interactions between users, such as text messages, comments, or loves, while for the location-based social network co-occurrences at events and regions were used. For the trading network edges were created when a user was buying an item from another user



- *Cosine similarity of product prices* We applied the same metric for product prices. Therefore, we graduated the prices by the following scheme: 0 – 5L\$, 6 – 10L\$, 11 – 20L\$, 21 – 50L\$, 51 – 200L\$, 201 – 500L\$, 501L\$ –  $\infty$ . We denoted the vectors with the number of products per price step for two users  $u$  and  $v$  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* We also calculated the cosine similarity for the user ratings of the products. Therefore, we classified the products in ten different rating schemes from .0 to 5.0 in incremental steps of .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 we computed the value of this feature as 
$$P_{RCos}(u, v) = \frac{\tau(u) \cdot \tau(v)}{\|\tau(u)\| \|\tau(v)\|}$$

Table 7 in this section gives a clear overview of the overall 57 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.

### 4.3 Learning methods and evaluation

As mentioned in the related work before, in the literature two different kinds of learning methods are typically employed to predict links in networks: supervised and unsupervised learning methods.

#### 4.3.1 Supervised learning

The first approach we employed to predict links between seller and buyers was a machine learning approach as also referred to as supervised learning. To do so, we created a balanced dataset of user pairs with and without purchases. Therefore, we used all 5376 user pairs that had a purchase relation in between. As negative samples we randomly chose the same amount of user pairs that had no trading interactions in between. 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 we 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 10,752 user pairs were split into a training set to determine characteristics of purchase interactions and a test set for verification with a tenfold cross-validation. This balanced sample of data resulted in a baseline of 50% for the trading prediction task when guessing at random.

We repeated all experiments ten times always choosing new random samples for the negative user pairs. As such, the results represent the averages of the respective values of the ten iterations. As a tool to run this task we chose the WEKA machine learning software (Hall et al. 2009).

As pointed out in the related work, machine learning strategies, such as decision trees, logistic regression, SVMs or other types of meta-learning strategies such as bagging or boosting are usually used for predicting links in social networks and they work remarkably well. As such we applied a series of methods available in the Weka

machine learning framework, such as Naive Bayes, Bayes Networks, Decision Tables, Logistic Regression, Random Forest, J48 Tree, and Random Tree. For meta-learning we employed bagging and boosting with the same classifiers and also tested stacking. Due to space limitations, we can only present the results of a handful of approaches. The approaches we selected were Random Forest, Logistic Regression and Naive Bayes. They showed not only the best results of the classic learning methods investigated but are also easy to implement in a real world system. The best overall meta-learning approach was bagging with Random Tree, which is also included.

As an evaluation metric Accuracy was chosen, as the positive and negative training examples are balanced.

### 4.3.2 Unsupervised learning

The second approach employed to predict trading interactions was an unsupervised learning approach in the form of a user-based collaborative filtering technique, as e.g., also proposed by Liben-Nowell and Kleinberg (2007). The intuition behind this idea was that buyers who are similar to each other will behave in a similar manner in the marketplace (Schafer et al. 2007).

We used the non-probabilistic user-based  $k$ -nearest neighbors algorithm (kNN), where for each buyer in the combined network we find their  $k$ -nearest neighbors, or more precisely, the  $k$  most similar buyers based on each individual feature and several feature sets. In a given data points collection a nearest neighbor of a query point is a data point that is closest to the query point (Beyer et al. 1999).  $k$  defines the size of the neighborhood, for example, for  $k = 10$  the 10 most similar buyers (to the given buyer) are considered based on the respective feature or feature set. Afterwards, we recommend the top- $N$  sellers who had a trading relation with the buyers computed via kNN. Finally, we compare these top- $N$  predicted sellers with the real sellers of the origin buyer.

In order to have a fair comparison between all features and feature sets, we first normalized all feature values. We applied various numbers for the parameters  $k$  and  $N$ . In this paper we only present the results of the parameters which performed best:  $k = 100$  and  $N = 5$ .

To evaluate this approach, we used the *mean average precision (MAP)* (Yue et al. 2007) as performance and correctness measure, and the *normalized discounted cumulative gain (nDCG)* (Yilmaz et al. 2008) as a measure for the ranking quality. We obtained the *MAP* by computing the mean over the *APs* from all buyers, defined as follows:

$$AP@n = \frac{\sum_{k=1}^n (P(k) \times rel(k))}{\#true\ sellers}.$$

We denoted *AP* as the average precision for a buyer with  $P(k)$  as the precision at cutoff  $k$  in the predicted sellers list.  $rel(k)$  is 1 if the  $k$ th seller in the list is predicted correctly, and 0 otherwise.

We compute *nDCG* as follow:

$$nDCG@n = \frac{DCG@n}{IDCG@n},$$

with

$$DCG@n = \sum_{k=1}^n \frac{rel(k)}{\log_2(k+1)}.$$

We assume that *IDCG* is the *DCG* in ideal ordering.

Additionally, we report the *user coverage (UC)* for every used feature and feature set to show for which fraction of users these were available.

#### 4.4 Statistical comparison: sellers versus buyers

To show the differences between user pairs with and without trading interactions we computed the mean values of the features by simply calculating the average of the feature values of the involved edges. We calculated the significance of each feature in several steps, similarly to Bischoff (2012): First, we computed the *Levene test*—introduced by Levene (1960)—with the positive and negative edges to test for equal variances. If the *p-value* of this function was below .01, we calculated the *Wilcoxon rank-sum test*, otherwise the *two-sided Kolmogorov-Smirnov test*. The returning *p-value* was the crucial measure for the significance of a feature. Since we randomly chose the negative user pairs ten times (see Sect. 4.3), we did this procedure also ten times and finally computed the mean of the values.

## 5 Results

This section presents the results of the implemented experiments. First, we show the mean values of the comparison between user pairs with and without trading interactions, the information gain of each feature and the Accuracy value with bagging, the *nDCG*, the *MAP*, and the *UC*, for each of the 57 features. Finally, we present a correlation heat map of all features and the trading prediction results for different feature combinations.

### 5.1 Individual features (RQ1)

In Table 7 we present the mean values and significances of all features of the three networks for user pairs with ( $\emptyset\text{Val}_w$ ) and without trading interactions ( $\emptyset\text{Val}_{wo}$ ). For the online social network we observed the highest significant differences for the preferential attachment score features  $O_{PS}^+$ ,  $O_{PS}^-$ . Although the values are tiny the Jaccard's coefficient of groups feature exhibits the most significant differences of all homophilic features of the online social network. For the location-based social network the features with the highest significant differences are the Katz features  $L_K$  and the common events feature  $E_C$ . For  $L_K$  user pairs with trading interactions have values about up to 38 times larger on average than user pairs without trading interactions. For  $E_C$  the differences between the user pairs are even higher. With values on average 50 times higher for user pairs with trading interactions this feature shows the most significant differences of all features used in this paper. For the topological features of the trading network we observed the largest differences between user pairs with and without trading interactions for the preferential attachment out feature  $T_{PS}^+$ . The cosine similarity of product categories  $P_{CCos}$  is the feature with the highest differences of all homophilic features of the trading network.

**Table 8** The Accuracy with bagging represents the quality of each feature for predicting trading interactions with supervised learning

Feature	Description	Supervised	Unsupervised		
		Accuracy	nDCG	MAP	UC (%)
<i>Online social network</i>					
Topological					
$O_{CN}^+$	Num. common outgoing neighbors	.5066	.0303	.0080	41.33
$O_{CN}^-$	Num. common incoming neighbors	.5067	.0274	.0085	44.49
$O_{JC}^+$	Outgoing Jaccard's coefficient	.5065	.0311	.0084	41.33
$O_{JC}^-$	Incoming Jaccard's coefficient	.5068	.0308	.0097	44.49
$O_{PS}^+$	Preferential attachment +−	.5441	.0468	.0101	66.52
$O_{PS}^-$	Preferential attachment −+	<b>.5705</b>	.0428	.0143	<b>76.41</b>
$O_R$	Reciprocity of user communication	.5091	<b>.0623</b>	<b>.0261</b>	10.46
$O_{AA}$	Adamic–Adar	.5059	.0274	.0086	36.12
$O_{K001}$	Katz ( $\beta = .001$ )	.5111	.0410	.0118	44.76
$O_{K01}$	Katz ( $\beta = .01$ )	.5109	.0406	.0117	44.76
$O_{K1}$	Katz ( $\beta = .1$ )	.5111	.0404	.0118	44.76
$O_{RPR01}$	Rooted PageRank ( $\alpha = .01$ )	.5123	.0601	.0157	47.06
$O_{RPR05}$	Rooted PageRank ( $\alpha = .05$ )	.5116	.0559	.0143	47.06
$O_{RPR15}$	Rooted PageRank ( $\alpha = .15$ )	.5114	.0510	.0140	47.04
$O_{RPR3}$	Rooted PageRank ( $\alpha = .3$ )	.5121	.0484	.0135	47.02
$O_{RPR5}$	Rooted PageRank ( $\alpha = .5$ )	.5116	.0415	.0123	46.91
Homophilic					
$G_C$	Num. common groups	<b>.5333</b>	<b>.0894</b>	<b>.0336</b>	<b>75.24</b>
$G_{JC}$	Jaccard's coefficient of groups	.5327	.0874	.0332	<b>75.24</b>
$I_C$	Num. common interests	.5034	.0552	.0113	30.37
$I_{JC}$	Jaccard's coefficient of interests	.5025	.0573	.0122	30.37
$OI$	Num. interactions	.5076	.0418	.0173	27.14
$RR_C$	Num. common check-ins	.5009	.0313	.0088	22.52
$RR_{JC}$	Jaccard's coefficient of check-ins	.5008	.0299	.0093	22.52
$RR_O$	Overlap of check-ins	.5008	.0297	.0093	22.52
$RF_C$	Num. common favored regions	.5139	.0564	.0196	73.54
$RF_{JC}$	Jaccard's coefficient of favored regions	.5141	.0554	.0201	73.54
$RF_O$	Overlap of favored regions	.5141	.0554	.0201	73.54
<i>Location-based social network</i>					
Topological					
$L_{CN}$	Num. common neighbors	.5206	.0656	.0233	95.33
$L_{JC}$	Jaccard's coefficient	.5202	.0641	.0227	95.33
$L_{AA}$	Adamic–Adar	.5170	.0638	.0227	95.33
$L_{DS}$	Num. days seen	.5171	.0408	.0177	74.74
$L_{MD}$	Mean distance	.5169	.0354	.0150	74.56
$L_{K001}$	Katz ( $\beta = .001$ )	.5159	<b>.0747</b>	.0187	63.88
$L_{K01}$	Katz ( $\beta = .01$ )	.5159	.0744	.0189	63.88
$L_{K1}$	Katz ( $\beta = .1$ )	.5140	.0714	.0183	63.88
$L_{RPR01}$	Rooted PageRank ( $\alpha = .01$ )	.5290	.0574	.0191	<b>96.96</b>
$L_{RPR05}$	Rooted PageRank ( $\alpha = .05$ )	.5317	.0668	.0234	<b>96.96</b>

**Table 8** (continued)

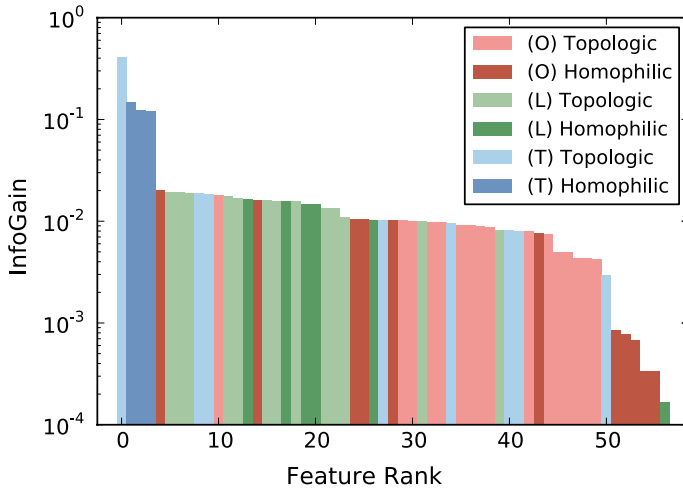
Feature	Description	Supervised	Unsupervised		
		Accuracy	nDCG	MAP	UC (%)
$L_{RPR15}$	Rooted PageRank ( $\alpha = .15$ )	<b>.5352</b>	.0717	.0241	<b>96.96</b>
$L_{RPR3}$	Rooted PageRank ( $\alpha = .3$ )	.5302	.0697	.0232	<b>96.96</b>
$L_{RPR5}$	Rooted PageRank ( $\alpha = .5$ )	.5270	.0721	<b>.0243</b>	<b>96.96</b>
Homophilic					
$E_C$	Num. common events	.5195	.0508	.0194	93.69
$E_{JC}$	Jaccard’s coefficient of events	.5195	.0531	.0202	93.69
$E_{CCos}$	Cosine similarity of event categories	.5271	.0575	.0168	94.93
$RE_C$	Num. common event regions	<b>.5346</b>	<b>.0687</b>	<b>.0239</b>	<b>99.66</b>
$RE_{JC}$	Jaccard’s coefficient of event regions	.5344	.0659	.0205	<b>99.66</b>
$RE_O$	Overlap of event regions	.5344	.0657	.0205	<b>99.66</b>
Trading network					
Topological					
$T_{CN}^+$	Num. common outgoing neighbors	.5101	.0598	.0099	10.26
$T_{CN}^-$	Num. common incoming neighbors	.5271	.9605	.7173	93.68
$T_{JC}^+$	Outgoing Jaccard’s coefficient	.5100	.0595	.0098	10.26
$T_{JC}^-$	Incoming Jaccard’s coefficient	.5271	.9660	<b>.7509</b>	93.68
$T_{PS}^+$	Preferential attachment +–	<b>.8881</b>	.1049	.0180	13.71
$T_{PS}^-$	Preferential attachment –+	.5351	.0271	.0057	<b>100</b>
$T_R$	Reciprocity of trading interactions	.5029	.2321	.0342	00.32
$T_{AA}$	Adamic–Adar	.5146	<b>.9676</b>	.5549	63.30
Homo.					
$P_{CCos}$	Cosine similarity of product categories	.7440	.2142	.0520	20.57
$P_{PCos}$	Cosine similarity of product prices	.7405	<b>.2266</b>	<b>.0553</b>	15.74
$P_{RCos}$	Cosine similarity of product ratings	<b>.7530</b>	.1792	.0375	<b>20.58</b>

nDCG, MAP, and UC show the predictive power of each feature with collaborative filtering. The best results in each feature set are highlighted in bold face

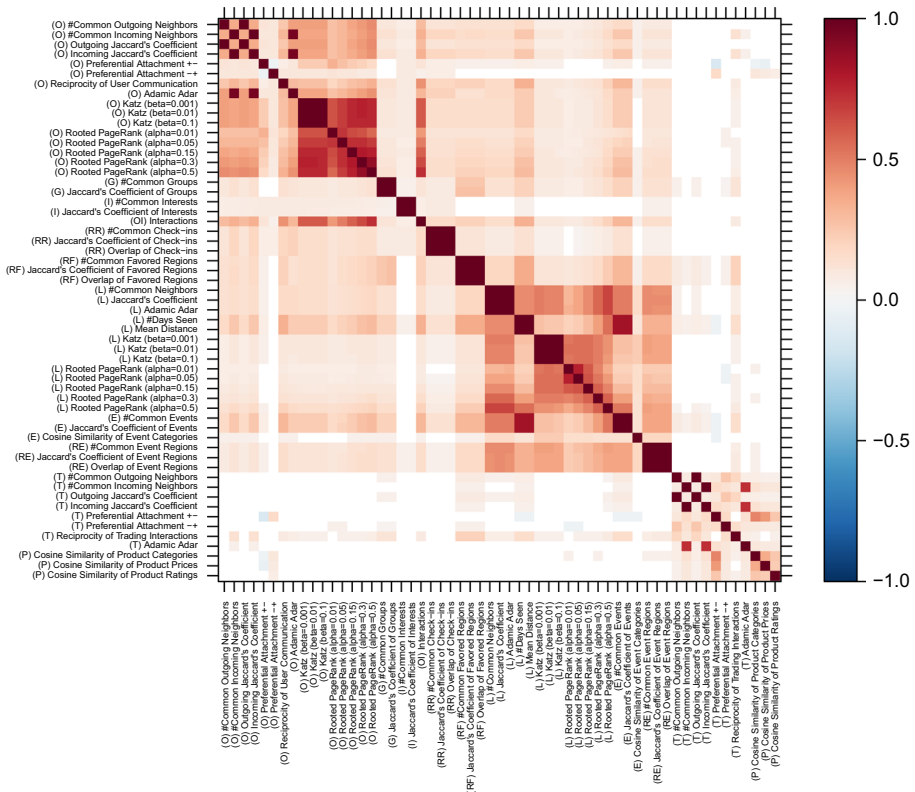
Table 8 provides an overview of the predictive power of each individual feature. The best performing topological online social network feature with the highest Accuracy and UC was the preferential attachment score  $O_{PS}^-$ . With unsupervised learning the best topological feature was the reciprocity of user communication  $O_R$ . The most valuable homophilic feature of the online social network was the number of common groups feature  $G_C$ . Although the result values are quite low, for the location-based social network the best performing features were the path-based measures rooted PageRank  $L_{RPR}$  and Katz  $L_K$ , and the number of common event regions  $RE_C$ . With the best Accuracy value of .8881 and the highest information gain, the preferential attachment feature  $T_{PS}^+$  of the trading network performed best.

Regarding unsupervised learning, the incoming Jaccard’s coefficient  $T_{JC}^-$  had the highest MAP and the Adamic–Adar measure  $T_{AA}$  had the best nDCG. Considering only homophilic features of the trading network, all three cosine similarity features performed quite similarly and were in the top four features regarding the information gain.

Figure 1 provides further insights to the question which feature is the most useful one employing Information Gain rank correlation analysis on all features. The colors of the



**Fig. 1** Quality ranking of the features according to their Information Gain. The topologic and homophilic feature sets of the online social network (*O*), the location-based social network (*L*), and the trading network (*T*) are color-coded. The most useful features are related to the trading network



**Fig. 2** A heat map indicating the Spearman feature cross-correlation values and showing the significant ( $p < 0.001$ ) correlations of all 57 features

bars indicate the feature group. The features with the highest information gain are again the ones from the trading network followed by location-based features.

### 5.2 Feature sets (RQ2–RQ3)

In predictive modeling when combining features to sets of features, it is common to take off with a correlation analysis. A correlation analysis typically helps in understanding better whether there are multi-collinearity issues which may create a problem or not later when features are combined. The correlation heat map in Fig. 2 shows the correlations between all 57 features at a significance level at  $p < 0.001$ . For several attributes, we obtained high correlations between the Jaccard’s coefficient, the Adamic–Adar, the observation and the common sets for the respective attribute. Apart from these expected findings, there are high correlations between the path measures Katz and rooted PageRank in the online social network. Also the interactions feature correlate with the path measures. As expected, the features about the number of days two users have met each other, and the mean distance between two users correlate with the events features, since they have similar data bases. The idea in this paper to use additional data sources instead of just trading network features to improve the trading predictions gets strengthened, since there are no significant correlations between the trading network features and features from the other data sources.

Table 9 provides an overview of how the several feature sets of all used data sources performed predicting trading interactions. We analyzed each feature set on its own. Using only online social network features we attained reasonable Accuracy values of up to .6275. A bit worse performed the combination of location-based social network features with a

**Table 9** The Accuracy values with random forest, logistic regression, naive bayes, and bagging show the predictive power of several feature sets

Feature set	Rand. For.	Logistic	Naive bay.	Bag-ging	nDCG	MAP	UC (%)	Num. features
<i>Single Networks</i>								
Online social (Homo)	.5380	.5424	.5397	.5439	.0790	.0282	91.73	11
Online social (Topo)	.5941	.5974	.5183	.6065	.0403	.0127	100	16
Online social (All)	.6106	.6157	.5476	.6275	.0756	.0263	100	27
Location-based (Homo)	.5458	.5363	.5367	.5555	.0525	.0161	99.92	6
Location-based (Topo)	.5361	.5250	.5273	.5400	.0580	.0205	96.96	13
Location-based (All)	.5460	.5406	.5408	.5544	.0509	.0155	99.97	19
Trading (Homo)	.7827	.6981	.7033	.8103	.1567	.0355	20.58	3
Trading (Topo)	.8769	<b>.8935</b>	<b>.7722</b>	.8920	<b>.8537</b>	<b>.6200</b>	100	8
Trading (All)	<b>.8963</b>	.8694	.7652	<b>.9119</b>	.6797	.5283	100	11
<i>Combined networks</i>								
Online + Location	.6006	.5988	.5556	.6267	.0518	.0147	100	46
Online + Trading	.9065	.8869	<b>.7547</b>	.9233	<b>.6440</b>	<b>.4507</b>	100	38
Location + Trading	.9032	.8834	.7437	.9211	.1306	.0794	100	30
Online + Location + Trading	<b>.9073</b>	<b>.8896</b>	.6998	<b>.9248</b>	.2493	.1374	100	57

nDCG, MAP, and UC represent the results for the unsupervised learning approach. Best results in each set are highlighted in bold face

Accuracy values of up to .5555. Combining the features of these two networks could not result in a performance boost, since the features of the online social network on its own are apparently quite tough for the seller buyer prediction task. We obtained the best prediction results using features of the trading network. Using only trading network features resulted in Accuracy values up to .9119. This astonishing result can be explained with the minor advantage the trading network has, because the trading interactions we tried to predict in our experiments originate from this network. Adding the online or the location-based social network features or both to the trading network features could slightly increase the prediction result about  $\approx 1.5\%$ .

Moreover, we observed that our topological features are more suitable than our homophilic features for the prediction of trading interactions as Table 9 also shows. This means that, by utilizing information of a network structure better trading prediction results could be obtained than by exploiting homophilic features regarding attributes of the users. The results of the collaborative filtering approach substantiate this finding.

## 6 Summary and discussion

The main findings with respect to our RQs can be summarized as follows:

- *RQ1* As expected, the best trading prediction results were obtained employing features from the trading network. The preferential attachment score with an Accuracy value of .8881 and the highest information gain was the best performing feature overall by far. It also exhibited the large significant differences between user pairs with and without trading interactions. Features of the online social network and the location-based network were also to some extent useful and could achieve Accuracy values up .5705 and .5352. The best individual features here were the Preferential Attachment feature in the social network and the rooted PageRank feature in the location-based network.
- *RQ2* In general, the results of our experiments show that topological features are more suitable than homophilic features for the prediction of trading interactions, since the Accuracy values of the topological feature sets were crucial higher than the values of the homophilic feature sets. This means that, for trading predictions the network structure is more useful than other user related attributes represented through our homophilic features. The location-based social network feature set obtained the worst prediction results of the three network sources with an Accuracy value with bagging of .5544. To exceed the “border” of a 60% prediction probability, it was necessary to use online social network features (.6275) or combine online and location-based social network features (.6267). Since the trading network has a minor advantage, because the trading interactions we tried to predict in our experiments originate from this network, the best results with Accuracy values up to .9119 could be achieved with the trading network feature sets. Adding online and/or location-based social network feature sets to them could slightly increase the prediction probability up to .9248. Conclusively, it could be said that online and location-based social network information on their own or in combination could result trading interaction prediction performance at an acceptable scale, which is convenient in specific setting, e.g., for cold-start prediction settings. As expected, the trading network information improved prediction performance strongly; it does not necessarily require the addition of further information of other network sources for trading interaction predictions. However, as noted above, trading informa-



tion may not be available in all cases, therefore, the results show when the other networks can then compensate here.

- *RQ3* Finally, we performed extensive experiments using both unsupervised as well as supervised learning approaches. As expected, the supervised approaches were able to score quite well, especially utilizing an ensemble classifier (bagging). Also, unsupervised strategies, which are more suitable in certain application settings where no ground truth information is available scored sufficiently.

Altogether, our results show important implications concerning personalization and recommendation approaches, e.g., Eirinaki et al. (2018). Regarding business and managerial applications, as already has been laid out in first fundamental investigations on buyer seller networks, e.g., Thorelli (1986) and Kranton and Minehart (2001), recommender systems in those areas play a decisive role in e-commerce, e.g., Schafer et al. (1999) and Linden et al. (2003). In particular, this specifically relates to the combined environments of digital as well as physical network structures (Liu et al. 2018), and how to effectively build recommender systems in those areas.

In the physical (real) world, for example, this relates to many systems which have a similar setup such as ebay<sup>10</sup> and other online stores, where there are important links and commonalities compared to Second Life as shown in Szell et al. (2012), Lehdonvirta (2009) and Guo et al. (2011). Then, the relations between the actors in these networks can be investigated from different (feature) perspectives—both from the physical as well as the online perspective; based on the results of this work—concerning the different feature sets and their impact classification approaches can be devised, making use of the available data in the best possible and cost-efficient way in order to optimize criteria such as predictive performance, recommendation diversity, or the available number of recommendations (Herlocker et al. 2004).

Finally, explainable recommenders, i.e., explanation-awareness (Atzmueller and Roth-Berghofer 2010; Nunes and Jannach 2017; Tintarev and Masthoff 2007) concerning the recommendations is very important, which is also enabled using the respective feature sets from multiple network perspectives. In addition, both supervised as well as unsupervised techniques can be applied here. In our experiments, random forest, logistic regression, naive bayes and bagging showed the best results which makes them very good candidates for providing explanations on the recommendations; because these are “white-box classifiers”, they can provide insights into the factors for inferring a certain recommendation based on the used parameters, and their weighting, e.g., Ribeiro et al. (2016), Li and Huan (2017) and Biran and Cotton (2017).

## 7 Conclusions and future work

In our work, we collected data from three different sources of Second Life—an online social network, a location-based social network and a trading network. Overall, we computed 57 topological and homophilic features to measure the similarities between user pairs and conducted several experiments predicting trading interactions.

---

<sup>10</sup> <http://ebay.com>.

In that way, we incorporated and analyzed the multiplex structure of the different networks and the individual features both from an individual as well as a collective perspective. This allowed us to identify the impact of the features from those networks both from detailed to aggregated level in order to derive actionable insights and implications from the analysis, e.g., for prediction, recommendation and marketing.

As already mentioned, this paper 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. When calculating a feature between two users about already traded products, for example, the products could be weighted in a way where the older trades would not be that important as newer ones. Furthermore, 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. Here, previous work that we have presented in Lacic et al. (2015) provides a good starting point for the mentioned future analysis directions.

A further interesting direction concerns the induction of other types of features such as recently proposed in Lee et al. (2016), e.g., employing Latent Dirichlet Allocation (LDA) (Blei et al. 2003) to find latent relations between users given their biographical information. However, more refined topic modeling approaches, such as proposed by Weng et al. (2010), would be needed in the scope of the Second Life dataset, since biographic information in the user profiles is rather sparse and typically less than 100 characters in length.

Another very interesting extension of this work would be to study the problem from a more theoretical/economical background. The study at hand does not do this extensively. Instead, we based our assumptions mostly on the link prediction and recommender systems research literature and common sociological factors indicating interactions. Building more grounded theoretical models would potentially also help to understand the nature of the problem better. So far, we have just a rough estimate. While our study suggests that predicting seller-buyer interactions from social network data is hard, Guo et al. (2011) estimates this as a moderate-hard predicting problem (at least their experimental setup and dataset suggests this). As such, more research on this problem in different kinds of datasets is needed. So far we can only claim that the problem at hand is rather easy to resolve in the Second Life dataset given trading interactions alone, while given location-based and online social network data alone, it is not.

Finally, since using the data of Second Life, the experiments in this paper were based on a virtual world. A rather important and relevant research direction concerns the understanding and modeling of digital and physical network structures and the behavior of actors therein. Therefore, an important task in the future could be to investigate how the experiments would perform if data of the “real” world combined with online social data were used.

## References

- Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3), 211–230.
- Al Hasan, M., & Zaki, M. J. (2011). A survey of link prediction in social networks. In *Social network data analytics* (pp. 243–275) Berlin: Springer.
- Atzmueller, M., & Roth-Berghofer, T. (2010). The mining and analysis continuum of explaining uncovered. In *Proceedings of 30th SGAI international conference on artificial intelligence (AI-2010)*.

- Backstrom, L., & Leskovec, J. (2011). Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining (WSDM)* (pp. 635–644).
- Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509–512.
- Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and Its Applications*, 311, 590–614.
- Barbieri, N., Bonchi, F., & Manco, G. (2014). Who to follow and why: Link prediction with explanations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM* (pp. 1266–1275).
- Beyer, K. S., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "Nearest Neighbor" meaningful? In *Proceedings of the 7th international conference on database theory* (pp. 217–235). Berlin: Springer.
- Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI*.
- Bischoff, K. (2012). We love rock 'n' roll: analyzing and predicting friendship links in Last.fm. In *Proceedings of the 3rd annual ACM web science conference, ACM* (pp. 47–56).
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1–7), 107–117.
- Chakrabarti, S. (2007). Dynamic personalized PageRank in entity-relation graphs. In *Proceedings of the 16th international conference on World Wide Web, ACM, New York, NY, WWW '07* (pp. 571–580).
- Cheng, J., Romero, D. M., Meeder, B., & Kleinberg, J. (2011). 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)* (pp. 49–56).
- Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, 94, 95–120.
- Cranshaw, J., Toch, E., Hong, J., Kittur, A., & Sadeh, N. (2010). Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing, ACM* (pp. 119–128).
- DiMaggio, P., & Louch, H. (1998). Socially embedded consumer transactions: For what kinds of purchases do people most often use networks? *American Sociological Review*, 63(5), 619–637.
- Eberhard, L., & Trattner, C. (2016). Recommending sellers to buyers in virtual marketplaces leveraging social information. In *Proceedings of the 25th international conference companion on World Wide Web, international World Wide Web conferences steering committee, republic and Canton of Geneva, Switzerland, WWW '16 Companion* (pp. 559–564).
- Eirinaki, M., Gao, J., Varlamis, I., & Tserpes, K. (2018). Recommender systems for large-scale social networks: A review of challenges and solutions. *Future Generation Computer Systems*, 78(1), 413–418.
- Fire, M., Tenenboim, L., Lesser, O., Puzis, R., Rokach, L., & Elovici, Y. (2011). 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), IEEE* (pp. 73–80).
- Fire, M., Tenenboim-Chekina, L., Puzis, R., Lesser, O., Rokach, L., & Elovici, Y. (2013). Computationally efficient link prediction in a variety of social networks. *ACM TIST*, 5(1), 10:1–10:25.
- Getoor, L., & Diehl, C. P. (2005). Link mining: A survey. *ACM Sigkdd Explorations Newsletter*, 7(2), 3–12.
- Guo, S., Wang, M., & Leskovec, J. (2011). The role of social networks in online shopping: information passing, price of trust, and consumer choice. In *ACM Conference on Electronic Commerce* (pp. 157–166).
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: An update. *IGKDD Explorations Newsletter*, 11(1), 10–18.
- Hasan, M. A., Chaoji, V., Salem, S., & Zaki, M. (2006). Link prediction using supervised learning. In *Proceedings of SDM 06 workshop on Link Analysis, Counterterrorism and Security*.
- Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1), 5–53.
- Jones, J. J., Settle, J. E., Bond, R. M., Fariss, C. J., Marlow, C., & Fowler, J. H. (2013). Inferring tie strength from online directed behavior. *PLoS ONE*, 8(1), e52168.
- Katz, L. (1953a). A new status index derived from sociometric analysis. *Psychometrika*, 18(1), 39–43.
- Katz, L. (1953b). A new status index derived from sociometric analysis. *Psychometrika*, 18(1), 39–43.

- Kibanov, M., Atzmueller, M., Illig, J., Scholz, C., Barrat, A., Cattuto, C., & Stumme, G. (2015). Is web content a good proxy for real-life interaction? A case study considering online and offline interactions of computer scientists. In *Proceedings of ASONAM Boston, MA*: IEEE Press.
- Kranton, R. E., & Minehart, D. F. (2001). A theory of buyer-seller networks. *American Economic Review*, *91*(3), 485–508.
- La, C. A., & Michiardi, P. (2008). Characterizing User Mobility in Second Life. In *Proceedings of the first workshop on Online social networks, ACM, WOSN '08*.
- Lacic, E., Kowald, D., Eberhard, L., Trattner, C., Parra, D., & Marinho, L. B. (2015). Utilizing online social network and location-based data to recommend products and categories in online marketplaces. In *Mining, modeling, and recommending 'Things' in social media* (pp. 96–115) Berlin: Springer.
- Lee, G. M., Qiu, L., & Whinston, A. B. (2016). A friend like me: modeling network formation in a location-based social network. *Journal of Management Information Systems*, *33*(4), 1008–1033.
- Lehdonvirta, V. (2009). Virtual item sales as a revenue model: identifying attributes that drive purchase decisions. *Electronic Commerce Research*, *9*(1), 97–113.
- Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010). Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World Wide Web, ACM* (pp. 641–650).
- Levene, H. (1960). In contributions to probability and statistics: Essays in honor of Harold Hotelling, Olkin, I., et al. eds, (pp. 278–292) Stanford : Stanford University Press.
- Li, X., & Huan, J. (2017). Constructivism learning: A learningparadigm for transparent predictive analytics. In *Proceedings of SIGKDD, ACM* (pp. 285–294).
- Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, *58*(7), 1019–1031.
- Liben-Nowell, D., & Kleinberg, J. M. (2003). The link prediction problem for social networks. In *CIKM* (pp. 556–559).
- Lichtenwalter, R., & Chawla, N. V. (2011). Lpmade: Link prediction made easy. *Journal of Machine Learning Research*, *12*, 2489–2492.
- Lichtenwalter, R., & Chawla, N. V. (2012). Vertex collocation profiles: Subgraph counting for link analysis and prediction. In *WWW* (pp. 1019–1028).
- Lichtenwalter, R., Lussier, J. T., & Chawla, N. V. (2010). New perspectives and methods in link prediction. In *KDD* (pp. 243–252).
- Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, *7*(1), 76–80.
- Liu, D. R., Chou, Y. C., Chung, C. C., & Liao, H. Y. (2018). Recommender system based on social influence and the virtual house bandwagon effect in virtual worlds. *Kybernetes*, *47*(3), 587–604.
- Lü, L., & Zhou, T. (2010). Link prediction in weighted networks: The role of weak ties. *EPL (Europhysics Letters)*, *89*(18), 001.
- Lu, Z., Savas, B., Tang, W., & Dhillon, I. S. (2010). Supervised link prediction using multiple sources. In *2010 IEEE 10th International Conference on Data Mining (ICDM)* (pp. 923–928). IEEE.
- Martínez, V., Berzal, F., & Cubero, J. C. (2016). A survey of link prediction in complex networks. *ACM Computing Surveys (CSUR)*, *49*(4), 69.
- McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444.
- Menon, A. K., & Elkan, C. (2011). Link prediction via matrix factorization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)* (pp. 437–452). Berlin: Springer.
- Murata, T., & Moriyasu, S. (2007). Link prediction of social networks based on weighted proximity measures. In *Web intelligence* (pp. 85–88).
- Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks. *Physical Review E*, *64*(2), 025,102.
- Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 1–52.
- Pearson, K. (1905). The problem of the random walk. *Nature*, *72*(1865), 294.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. <https://doi.org/10.1023/A:1022643204877>.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of ACM SIGKDD, ACM* (pp. 1135–1144).
- Ripley, R. M., Snijders, T. A., Boda, Z., Vörös, A., & Preciado, P. (2011). *Manual for RSIENA*. Oxford: Department of Statistics, Nuffield College 1, University of Oxford.

- Rowe, M., Stankovic, M., & Alani, H. (2012). 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, ISWC12* (pp. 476–491). New York: Springer.
- Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York, NY: McGraw-Hill Inc.
- Scellato, S., Noulas, A., & Mascolo, C. (2011). Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM* (pp. 1046–1054).
- Schafer, J., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In P. Brusilovsky, A. Kobsa, W. Nejdl (Eds.) *The adaptive web, lecture notes in computer science* (Vol. 4321, pp. 291–324). Heidelberg: Springer.
- Schafer, J. B., Konstan, J., & Riedl, J. (1999). Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on Electronic commerce, ACM* (pp. 158–166).
- Scholz, C., Atzmueller, M., & Stumme, G. (2012). On the predictability of human contacts: Influence factors and the strength of stronger ties. In *IEEE Computer Society SocialCom 2012*, Los Alamitos, CA, USA.
- Scholz, C., Atzmueller, M., Barrat, A., Cattuto, C., & Stumme, G. (2013a). New insights and methods for predicting face-to-face contacts. In *Proceedings of 7th International AAAI conference on weblogs and social media*.
- Scholz, C., Atzmueller, M., Kibanov, M., & Stumme, G. (2013b). How do people link? Analysis of contact structures in human face-to-face proximity networks. In *Proceedings of ASONAM 2013*, New York, NY: ACM Press.
- Scholz, C., Illig, J., Atzmueller, M., & Stumme, G. (2014). On the predictability of talk attendance at academic conferences. In *Proceedings of 25th ACM conference on hypertext and social media* New York, NY: ACM Press.
- Spitzer, F. (2013). *Principles of random walk* (Vol. 34). New York: Springer.
- Steurer, M., & Trattner, C. (2013a). Acquaintance or partner? Predicting partnership in online and location-based social networks. In *IEEE/ACM ASONAM* (pp. 1–8).
- Steurer, M., & Trattner, C. (2013b). Predicting interactions in online social networks: An experiment in second life. In *Proceedings of the 4th international workshop on modeling social media* (pp. 5:1–5:8). ACM.
- Steurer, M., & Trattner, C. (2013c). 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, Lecture Notes in Computer Science* (Vol. 8329, pp. 108–127). Berlin Heidelberg: Springer.
- Steurer, M., Trattner, C., & Helic, D. (2013). Predicting social interactions from different sources of location-based knowledge. In *The third international conference on social eco-informatics, Lisbon, Portugal, SOTICS 2013* (pp. 8–13).
- Szell, M., Sinatra, R., Petri, G., Thurner, S., & Latora, V. (2012). Understanding mobility in a social petri dish. *Scientific reports* 2.
- Thelwall, M. (2009). Homophily in MySpace. *Journal of the American Society for Information Science and Technology*, 60(2), 219–231.
- Thiele, L., Atzmueller, M., Stumme, G., & Kauffeld, S. (2018). Frequently and/or durable? The predictive impact of initial face-to-face contacts on the formation and evolution of developmental peer network relationships. *Psychology*.
- Thorelli, H. B. (1986). Networks: Between markets and hierarchies. *Strategic Management Journal*, 7(1), 37–51.
- Tintarev, N., & Masthoff, J. (2007). A survey of explanations in recommender systems. In *Data engineering workshop, 2007 IEEE 23rd International Conference on, IEEE* (pp. 801–810).
- Trattner, C., & Steurer, M. (2015). Detecting partnership in location-based and online social networks. *Social Network Analysis and Mining*, 5(1), 42.
- Varvello, M., & Voelker, G. M. (2010). 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, ACM, NOSSDAV '10* (pp. 9–14).
- Varvello, M., Picconi, F., Diot, C., & Biersack, E. (2008). Is there life in second life? In *Conext'08*, Madrid, Spain.
- Wang, P., Xu, B., Wu, Y., & Zhou, X. (2015). Link prediction in social networks: The State-of-the-Art. *Science China Information Sciences*, 58(1), 1–38.

- Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010). Twiterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining, ACM* (pp. 261–270).
- Yilmaz, E., Kanoulas, E., & Aslam, J. A. (2008). A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM* (pp. 603–610).
- Yue, Y., Finley, T., Radlinski, F., & Joachims, T. (2007). A support vector method for optimizing average precision. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, ACM, New York, NY, USA, SIGIR '07* (pp. 271–278).
- Zhang, J., & Philip, S. Y. (2014). *Link prediction across heterogeneous social networks: A survey*. Social networks.
- Zhang, J., Kong, X., & Philip, S. Y. (2013). Predicting social links for new users across aligned heterogeneous social networks. In *2013 IEEE 13th international conference on data mining (ICDM)* (pp. 1289–1294).
- Zhang, Y., & Pennacchiotti, M. (2013). Predicting purchase behaviors from social media. In *Proceedings of the 22nd international conference on World Wide Web, WWW '13* (pp. 1521–1532).
- Zhuang, H., Chin, A., Wu, S., Wang, W., Wang, X., & Tang, J. (2012a). Inferring geographic coincidence in ephemeral social networks. In: *ECML/PKDD Berlin*: Springer.
- Zhuang, H., Tang, J., Tang, W., Lou, T., Chin, A., & Wang, X. (2012b). Actively learning to infer social ties. *Data Mining Knowledge Discovery*, 25(2), 270–297.