Who will Interact with Whom? A Case-Study in Second Life using Online and Location-based Social Network Features to Predict Interactions between Users

Michael Steurer[1] and Christoph Trattner[2]

[1] IICM, Graz University of Technology
Inffeldgasse 16c
msteurer@iicm.edu
[2] Know-Center, Graz University of Technology
Inffeldgasse 13/5
ctrattner@know-center.at

Abstract. Although considerable amount of work has been conducted recently of how to predict links between users in online social media or networks, studies using features from different domains are rare. In this paper we present the latest results of a project that studies the extent to which interactions – in our case directed and bi-directed message communication – between users in online social networks can be predicted by looking at features obtained from online and location-based social network data. To that end, we conducted a number of experiments on data obtained from the virtual world of Second Life. As our results reveal, location-based social network features outperform online social network features if we try to predict interactions between users. However, if we try to predict, whether or not this communication was also reciprocal we find that online social network features seem to be superior.

Keywords: online social networks, location-based social networks, link prediction problem, predicting interactions, predicting reciprocity, virtual worlds, Second Life

1 Introduction

As a part of the recent hype on social network research, a high amount of attention and research activity was devoted to the problem of predicting links between users [17], *e.g.* the issue of forecasting whether or not two users u and v of a given online social network $G\langle V, E\rangle$ will interact with each other in the future. While considerable amount of work has been recently conducted of how to predict links between users in online social media or networks, studies utilizing information from domains are rare.

To contribute to this research, we present in this paper the latest results of a research project that aims to study the extent to which interactions – in our

case directed and bi-directed message communications –in online social networks can be predicted inducing features from online social network and location-based social network data. To tackle this issue we trained a binary classifier that learned the relations between users u and v based on a number of features induced from online social network and location-based social network data. For the purpose of our study we furthermore differentiated between two types of feature sets – network topological features and homophilic features [22]. Since it is nearly impossible to obtain rich large-scale real-world online social and location-based data, our investigation focused on the virtual world of Second Life, where we could easily find and mine both sources of data. We obtained data from a resource called *My Second Life* which is a large-scale online social network for residents of Second Life. This social network can be compared to Facebook but aims at a different target group: residents of Second Life who interact with each other by sharing text messages, comments, and loves. Additionally, we were able to collect location-based social network data of residents in the virtual world by implementing so-called in-world bots.

Overall, it is our interest to answer the following research questions:

- *RQ1:* To what extent do user pairs – interacting or not interacting with each other – differ based on social proximity features induced from the online social network and the location-based social network?

- *RQ2:* To what extent can we predict interactions between users and reciprocity of these interactions inducing features from both domains?

- *RQ3:* Which feature set (homophilic or topological) is most suitable to predict interactions between users and the reciprocity of these interactions.

To that end, we conducted a number of experiments using statistical methods and supervised learning approaches. As our statistical analysis reveals, there are many significant differences between user pairs with interactions and user pairs without interactions. For instance, users with an interactions on the online social network have a shorter average distance between them in the location-based social network. To predict these interactions with supervised learning, we find that location-based social network features outperform online social network features to a great extent. However, if we try to predict reciprocal message communication between users, online social network features seem to be superior. Finally, we find that there are no clear patterns whether or not homophilic or network topological features perform better to predict interactions or reciprocity between users.

All over the paper is structured as follows: In Section 2, we discuss related work. In Section 3 we shortly introduce the dataset used for our experiments. In Section 4 we outline the set of features used for our experiments in Section 5. Section 6 presents the results of our study. Finally, Section 7 discusses our findings and concludes the paper.

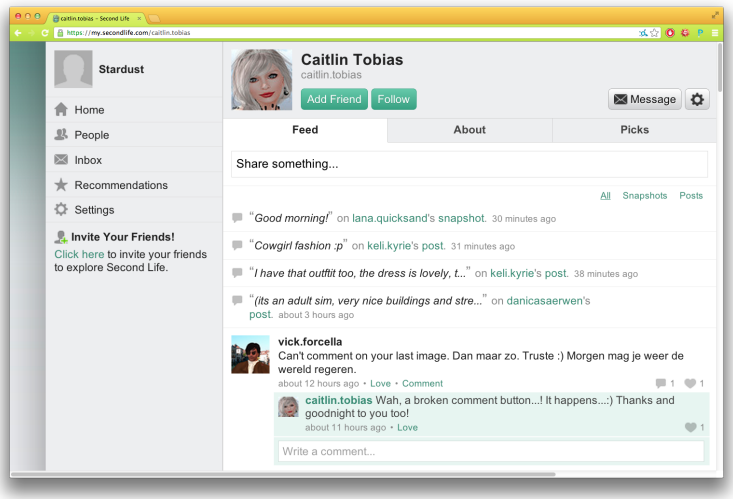

Fig. 1. Sample of a user profile in the online social network *My Second Life*. Users can *post* text message on their wall or can communicate with each other by *commenting* or *loving* onto each other's posts.

2 Related Work

Although considerable amount of work has been recently conducted of how to predict links between users in online social media, studies exploiting different kinds of knowledge sources for the link prediction problem are rare. An example is a study conducted by Cranshaw *et al.* where the authors collected location data and Facebook friendship data through a mobile app [6]. Based on a number of experiments they show that the so-called place-entropy features are best suited to predict friendship between users in Facebook. Interestingly and contrary to our study, Cranshaw *et al.* only looked at the mobile side, i.e. they did not investigate features induced directly from the social network. Furthermore, they only considered friendship links and did not look at communication links as we do in our study. Another related work in this context are the studies of Guy *et al.* [11], [12], [10] where the authors investigate the similarity between users exploiting 9 different sources of data classified into three different classes: *people*, *things*, and *places*. Looking at only semantic features such as tags, they find that the so-called "tagged-with" feature performs well in all three different data category sources.

Probably one of the first projects investigating the link prediction problem from the network topological perspective in the context of online social media is a work conducted by Golder and Yardi [8]. In their paper they study the micro-blogging service Twitter and find "that two structural characteristics, transitivity and mutuality, are significant predictors of the desire to form new ties". The first paper investigating the extent to which reciprocity could be predicted in the

online social media is a recent paper by Cheng *et al.* [4]. By applying a rich set of network based features including link prediction features from [17], they show that the so-called out-degree measure of a user in Twitter is the best feature to predict reciprocity. Another interesting work in this context is a study conducted by Yin *et al.* [23]. In their paper they investigate the link prediction problem within the micro-blogging system Twitter. The main contribution, apart from studying the performance of well established link prediction methods, is the introduction of a "novel personalized structure-based link prediction model" which "noticeably outperforms the state-of-the-art" methods. The first work studying the computational efficiency of network topological features in the online domain is a paper written by Fire *et al.* [7]. In their work they apply a rich set of over 20 features on a set of 5 different online social network sites with respect to their computational efficiency. Their study reveals that the so-called friends measure shows a good trade-off between accuracy and computational efficiency.

Another study in this context is a recent study conducted by Rowe *et al.* In their work [19] they study the link prediction problem, or the question who will follow whom, in the micro-blogging system *Tencent Weibo*. Looking at both – semantic and network topological features – they show that the predictability of links can be significantly improved by training a classifier that uses both. Although the work of Rowe *et al.* has considerable amount of overlap with our own work, their study only looked at features which could be directly induced from the online media site Tencent Weibo. Hence, contrary to our own work they did not include external knowledge such as location-based social network data as we do in our study. Finally, the last study to be mentioned is a work conducted by Scellato *et al.* [20]. Similar to our work they tried to exploit features from the location-based social network of *Gowalla* to predict links between users. However, in contrast to our work, they only focused on location-based social data and did not combine online social network and location-based social network data as we do in this paper. In their analysis over a period of three months they found that most of the links are formed between users that visit the same places or places that share similar properties.

3 Datasets

As stated in the introductory part of this paper we conducted our experiments on two types of datasets – online social network and location-based social data – both obtained from the virtual world of Second Life. The reasons for choosing Second Life over other real world sources are manifold: First, in contrast to networks such as Facebook, the online social network *My Second Life* does not restrict extensive crawling of user profiles. Second and contrary to real world online social networks, most profiles in *My Second Life* are public, i.e. we can mine a large fraction of the network. Third, in virtual worlds the location information of users can be harvested in an automated way whereas it is nearly impossible to obtain large-scale tracking data of users in the real world. In this section we describe the collection process for the data as used in our experiments.

3.1 Location-based Social Network Dataset

The collection of the location-based social network dataset in Second Life was a two stage process: First a list of popular locations from the Second Life Event calendar[3] was crawled. Second, overall 15 in-world agents so-called in-world-bots were implemented to teleport to these locations and gather location information of the users at place.

In detail the procedure was the following: In order to harvest all events in Second Life we implemented a Web-crawler that runs on a daily bases to obtain all publicly announced events on the Second Life Event calendar. Allover, we were able to obtain data of 218,245 unique events during a period of ten months starting in March 2012.

In order to collect location data of the users we implemented overall 15 in-world agents on the basis of the open source command-line client *libopenmetaverse*[4]. Due to the modularity of the tool, we were able to enhance the functionality of our agents to teleport around in the virtual world to collect location data of all surrounding users in a region. This location information comprised the current region, x and y coordinates of the location within this region, and a time stamp. The pool of agents was controlled by a centralized instance sending our in-world bots to ongoing events. Due to the large amount of concurrent events in several regions of Second Life and the constraint that a bot was only able to obtain data of one single region at the same time, our sampling rate was set to a limit of 15 minutes. All in all, we were able to obtain over 13 Million data samples of 190,160 unique users visiting events with this kind of approach [21].

3.2 Online Social Network Dataset

In July 2011 Linden Labs introduced an online social network called *My Second Life*[5] similar to other social networks such as Google+ or Facebook. Residents of the virtual world can log-in with their in-world credentials, access their friend lists and have a so-called *Feed* that can be compared to the Google+ Stream or the Facebook Wall. The social interaction with other users is done by sharing text messages, screenshots, comments and so-called loves which can be seen equally to a Like on Facebook or a Plus in Google+ (see Figure 1). Furthermore, users can enhance their profiles by adding personal information such as interests, groups, etc.

We attempted to download the profile data of all 190,160 users found by the avatar-bots. In the next step we parsed the interaction-partners of the these users and downloaded the profile information of the missing ones. This procedure was repeated until no new users could be found by our crawler anymore. Finally, this yielded in a dataset of 311,959 users with 300,657 of them opened to the public, and 135,181 with interactions on their feed.

[3] `http://secondlife.com/community/events/`
[4] `http://lib.openmetaverse.org/`
[5] `https://my.secondlife.com/`

Table 1. Basic metrics of the two networks and their combination used for the experiments.

Name	Location-based G_M	Online G_F	$G_{FM} = G_F + G_M$
Type	undirected	directed	directed
Nodes	131,349	135,181	37,118
Edges	2,343,683	209,653	1,043,172
Degree	35.7	3.1	56.2

4 Feature Sets

As already outlined, it is our interest to predict interactions between users in online social networks based on features induced from online social network and location-based social network data. To that end, we induced two different types of feature sets from our data sources: network topological and homophilic features [22]. In order to start with the description of the different features calculated for our experiments we first describe the networks derived from the collected data.

The first network, referred to as *online social network*, was based on data obtained from the users profile where every edge in this directed network indicates communication between two users. This yielded in a network with 135,181 users and 209,653 edges. The second network, referred to as *location-based social network*, was based on the users location data where every edge in this undirected network indicated that two users were seen concurrently in the same region on two different days. This yielded in a network with 142,507 nodes and 3,773,316 edges. A summery of both networks can be found in Table 1 and Figure 2 shows the degree distribution of the social network and location-based social network. Both networks show power-law qualities with an alpha of 1.55 and a corresponding fitting error of 0.13 for the online social network and and alpha value of 2.67 and a fitting error of 0.16 for the location-based social network [5].

4.1 Online Social Network: Topological Features

In social networks such as Facebook or Google+ the friendship of users is based on a mutual agreement where both confirm each other. In contrast to this, users of the online social network *My Second Life* can post onto each others' walls without this mutual agreement. Hence, as a consequence, we considered the social network as a directed graph $G_F\langle V_F, E_F\rangle$ with V_F representing the users and $e = (u, v) \in E_F$ if user u posted, commented, or liked something on the feed of user v.

We defined the set of the neighbors of a node $v \in G_F$ as $\Gamma(v) = \{u \mid (u, v) \in E_F \text{ or } (v, u) \in E_F\}$ and based on this definition of neighborhood we used the following topological features:

– *Common Neighbors* $F_{CN}(u, v)$. This represented number of interaction-partners two users had in common.

$$F_{CN}(u, v) = |\Gamma(u) \cap \Gamma(v)|$$

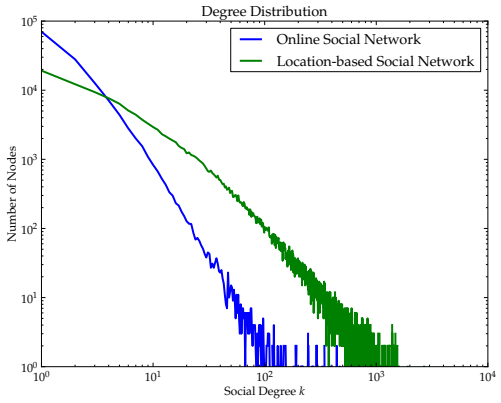

Fig. 2. Degree distributions for the online and the location-based social network.

For a directed network we split this into the number of common users $F^{+}_{CN}(u,v) = |\Gamma^{+}(u) \cap \Gamma^{+}(v)|$ to which both users sent messages to and the number of users $F^{-}_{CN}(u,v) = |\Gamma^{-}(u) \cap \Gamma^{-}(v)|$ from which both users received messages.

– *Jaccard's Coefficient* $F_{JC}(u,v)$. The ratio of the total number of neighbors and the number of common neighbors of two users was taken from [15] and is defined as follows.

$$F_{JC}(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$$

For directed networks this could be split into two coefficients for received messages $F^{-}_{JC}(u,v) = \frac{|\Gamma^{-}(u) \cap \Gamma^{-}(v)|}{|\Gamma^{-}(u) \cup \Gamma^{-}(v)|}$ and sent messages $F^{+}_{JC}(u,v) = \frac{|\Gamma^{+}(u) \cap \Gamma^{+}(v)|}{|\Gamma^{+}(u) \cup \Gamma^{+}(v)|}$.

– *Adamic Adar* $F_{AA}(u,v)$. Instead of just counting the number of common neighbors with Jaccard's Coefficient in a network, this feature adds weights to all neighbors of a pair of users [1].

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

According to Cheng *et al.* this can be transformed into $F^{-}_{AA}(u,v) = \sum_{z \in \Gamma^{-}(u) \cap \Gamma^{-}(v)} \frac{1}{log(|\Gamma^{-}(z)|)}$ for directed networks [4].

– *Preferential Attachment Score* $F_{PS}(u,v)$. This feature took into account that active users, i.e. users with many interaction partners, are more likely to form new relationships than users with not so many interactions [2].

$$F_{PS}(u,v) = |\Gamma(u)| \cdot |\Gamma(v)|$$

The score was applied to a directed network with two different features: $F^{+}_{PS}(u,v) = |\Gamma^{+}(u)| \cdot |\Gamma^{-}(v)|$, respectively $F^{-}_{PS}(u,v) = |\Gamma^{-}(u)| \cdot |\Gamma^{+}(v)|$ [4].

4.2 Online Social Network: Homophilic Features

As stated before, users in Second Life can enhance their online social network profile by adding additional meta-data information such as interests or groups. As observed by a number of previous studies in this area [19], [22], homophily is an important variable in the context of the link prediction problem. To account for factor, we defined a set of homophilic features which we calculated as group and interest similarity between users u, v. Formally, we defined the groups of a user u as $\Delta(u)$, respectively her interests as $\Psi(u)$.

- *Common Groups* $G_C(u,v)$. The first feature we induce is the so-called common groups measure. It is calculated as follows.

$$G_C(u,v) = |\Delta(u) \cap \Delta(v)|$$

- *Jaccard's Coefficient for Groups* $G_{JC}(u,v)$. The second feature, is the so-called Jaccard's coefficient for groups. It was calculated in the following form.

$$G_{JC}(u,v) = \frac{|\Delta(u) \cap \Delta(v)|}{|\Delta(u) \cup \Delta(v)|}$$

- *Common Interests* $I_C(u,v)$. The third homophilic feature, was the number of interests two users had in common.

$$I_C(u,v) = |\Psi(u) \cap \Psi(v)|$$

- *Jaccard's Coefficient for Interests* $I_{JC}(u,v)$. And finally the last feature, which is a combination of total interests and common interests of the users.

$$I_{JC}(u,v) = \frac{|\Psi(u) \cap \Psi(v)|}{|\Psi(u) \cup \Psi(v)|}$$

4.3 Location-based Social Network: Topological Features

We applied the same network topological feature calculations to the location-based social network as we did for the online social network. The network had edges between users that met on at least two days. Using this relation between in-world users defined the topological features similar to Section 4.1. Here, the neighbors of a node in the undirected location-based social network $G_M\langle V_M, E_M\rangle$ were defined as $\Theta(u) = \{v \mid (u,v) \in G_M\}$ and starting with this we defined the topological features as follows.

- *Common Neighbors* $M_{CN}(u,v)$.

$$M_{CN}(u,v) = |\Theta(u) \cap \Theta(v)|$$

– *Jaccard's Coefficient* $M_{JC}(u, v)$.

$$M_{JC}(u, v) = \frac{|\Theta(u) \cap \Theta(v)|}{|\Theta(u) \cup \Theta(v)|}$$

– *Adamic Adar* $M_{AA}(u, v)$.

$$M_{AA}(u, v) = \sum_{z \in \Theta(u) \cap \Theta(v)} \frac{1}{log(|\Theta(z)|)}$$

– *Preferential Attachment Score* $M_{PS}(u, v)$.

$$M_{PS}(u, v) = |\Theta(u)| \cdot |\Theta(v)|$$

4.4 Location-based Social Network: Homophilic Features

These features were based on the actual distance between users, the regions they visit, and the number of days where they co-occurred concurrently. Let $O(u, v)$ be the co-locations of user u and user v, when both users resided in the same region concurrently. An observation $o \in O(u, v)$ was 4-tuple of region r, time stamp t, location coordinates of user u: $l_u = (x_u, y_u)$ and user v: $l_v = (x_v, y_v)$.

– *Physical Distance* $A_D(u, v)$. Whenever two users were observed concurrently, we measured the distance between them based on their x and y coordinates. As a indicator for their overall physical closeness, we therefore computed the average physical Euclidean distance between two users for all observations where both were present in the same region concurrently.

$$A_D(u, v) = \frac{1}{|O(u, v)|} \sum_{o \in O(u,v)} \|o(l_u) - o(l_v)\|$$

– *Days Seen* $A_S(u, v)$. This feature indicated the number of days when two users have been observed in the same region concurrently.

The regions of a user were defined as $P(u) = \{\rho \in P \mid \text{user } u \text{ was observed in region } P\}$ and so we computed the region properties of users as follows:

– *Common Regions* $R_C(u, v)$. The number of regions two users visited, not necessarily at the same time.

$$R_C(u, v) = |P(u) \cap P(v)|$$

– *Regions Seen Concurrently* $R_S(u, v)$. In contrast to the Common Regions feature, this feature took only the regions into account where both users were observed in the same region concurrently.

– *Observations Together* $R_O(u, v)$. This feature was taken from Cranshaw *et al.* [6] and represented the number of total regions of two users divided by the sum of each user's number of regions.

$$R_O(u, v) = \frac{|P_u \cup P_v|}{|P_u| + |P_v|}$$

5 Experimental Setup

All in all, we conducted two different experiments to study the extent to which interactions between users in online social networks can be predicted. Both experiments were based on the combination of the *online social network* G_F and the *location-based social network* G_M described in Section 4. To that end, we followed the approach of Guha *et al.* [9] in both experiments who suggest to create two datasets with an equal number of "positive edges" and "negative edges" for the binary classification problem. This results in balanced datasets for the test- and the training data and therefore in a baseline of 50% for the prediction when guessing randomly. For the evaluation of the binary classification problem we employed different supervised learning algorithms and used the area under the ROC curve (AUC) as our main evaluation metric to determine the performance of our features [14], [18]. We justified our findings with a 10-fold cross validation approach using the WEKA machine-learning suite [13].

In this section we describe in detail how the trainings and test data set for both experiments was generated.

5.1 Predicting Interactions

The task here is to predict whether or not two users interacted with each other on the feed by using topological and homophilic information of the online social network and the location-based social network. In the first step we computed the edge-features for the user-pairs as described in Section 4 for both networks independently. Then, in the second step we created the intersection of both networks as directed graph $G_{FM}\langle V_{FM}, E_{FM}\rangle$ where $V_{FM} = \{v|v \in V_F, v \in V_M\}$, and $E_{FM} = \{(u,v)|(u,v) \in E_M, (u,v) \in E_F, v \text{ and } u \in V_{FM}\}$. This newly created network consisted of 37,118 nodes and 1,014,352 pairs with location co-occurrences ($(u,v) \in E_M$), 36,213 pairs with social interaction ($(u,v) \in E_F$), and 7,393 edges with both ($(u,v) \in E_M, E_F$).

For the binary classification problem we uniformly selected 2,500 user-pairs with a social interaction and a location co-occurrence ("positive edges") $\{e^+ = (u,v)|e^+ \in E_{FM},\ e^+ \in E_F,\ e^+ \in E_M\}$ and 2,500 user-pairs with a location co-occurrence but without a social interaction ("negative edges") $\{e^- = (u,v)|\ e^- \notin E_F,\ e^- \in E_M\}$. These edges, i.e. pairs of users, and the according edge features from both domains were used as datasets for all further evaluations and experiments.

5.2 Predicting Reciprocity

The task here is to predict whether two users had mutual activities on each other's wall, i.e. reciprocal interactions, by exploiting topological and homophilic information of the online social network and the location-based social network. We defined a reciprocal edge as $e'' = (u,v)|(u,v) \in G_F, (v,u) \in G_F$, a non-reciprocal edge as $e' = (u,v)|(u,v) \in G_F$, $(v,u) \notin G_F$, and used this difference for the binary classification problem. In contrast to the previous experiment we

Table 2. Means and standard errors of the features in the online social network and the location-based social network for the group of users having interactions with each other vs. the groups of users having no interactions (***=significant at level 0.001) .

	Features	Have Interactions	Have No Interactions
Online Social Network	Common Neighbors (in) $F^{-}_{CN}(u,v)^{***}$	2.81 ± 0.32	0.02 ± 0.00
	Common Neighbors (out) $F^{+}_{CN}(u,v)^{***}$	2.39 ± 0.27	0.01 ± 0.00
	Adamic Adar $F_{AA}(u,v)^{***}$	14.65 ± 1.28	1.71 ± 0.18
	Jaccard's Coefficient (in) $F^{-}_{JC}(u,v)^{***}$	0.05 ± 0.00	0.00 ± 0.00
	Jaccard's Coefficient (out) $F^{+}_{JC}(u,v)^{***}$	0.04 ± 0.00	0.00 ± 0.00
	Preferential Attachment (in) $F^{-}_{PS}(u,v)^{***}$	1566.55 ± 239.31	3.88 ± 0.64
	Preferential Attachment (out) $F^{+}_{PS}(u,v)^{***}$	2088.94 ± 441.14	4.92 ± 1.53
	Common Groups $G_C(u,v)^{***}$	1.92 ± 0.07	0.40 ± 0.02
	Jaccard's Coefficient $G_{JC}(u,v)^{***}$	0.05 ± 0.00	0.01 ± 0.00
	Common Interests $I_C(u,v)$	0.07 ± 0.01	0.02 ± 0.00
	Jaccard's Coefficient $I_{JC}(u,v)$	0.00 ± 0.00	0.00 ± 0.00
Location-based Social Network	Common Neighbors $M_{CN}(u,v)^{***}$	52.48 ± 4.98	83.61 ± 2.31
	Jaccard's Coefficient $M_{JC}(u,v)^{***}$	0.20 ± 0.00	0.10 ± 0.00
	Preferential Attachment $M_{PS}(u,v)^{***}$	218341.22 ± 164510.35	530640.88 ± 50352.29
	Adamic Adar $M_{AA}(u,v)^{***}$	26.89 ± 3.19	36.43 ± 0.98
	Regions Seen $R_S(u,v)^{***}$	2.81 ± 0.09	1.41 ± 0.02
	Common Regions $R_C(u,v)^{***}$	3.59 ± 0.34	3.03 ± 0.08
	Observations Together $R_O(u,v)^{***}$	0.22 ± 0.00	0.10 ± 0.00
	Distance $A_D(u,v)^{***}$	10.32 ± 0.36	38.13 ± 0.95
	Days Seen $A_S(u,v)^{***}$	7.34 ± 0.21	3.98 ± 0.09

considered the online social network as undirected network for the computation of the edge-features but retained information about the reciprocity of the interactions. The edge features for the location-based social network were again considered as undirected. For the actual experiment we combined the online social network and the location-based social network to a new undirected network referred to as $G'_{FM}\langle V'_{FM}, E'_{FM}\rangle$ where $V'_{FM} = \{v | v \in V_F, v \in V_M\}$, and $E'_{FM} = \{(u,v) | (u,v) \in E_M, (u,v) \in E_F \text{ or } (v,u) \in E_F, v \text{ and } u \in V'_{FM}\}$. Out of the 7,393 user-pairs with a social interaction and a location co-occurrence we identified 1,431 reciprocal edges and 4,531 non-reciprocal edges in the online social network. For the binary classification task we uniformly selected pairs of users from the undirected network G'_{FM} with 1,000 reciprocal edges ("positive edges") and non-reciprocal edges ("negative edges") each. These edges, i.e. user-pairs with the according features, were used for all further evaluations and experiments.

6 Results

Before we start with the analysis of how to predict interactions between users, we show the differences between user pairs with and without interactions in the social network, respectively user pairs with reciprocal and non-reciprocal interactions for both domains. Both the Anderson-Darling test and the one-sampled Kolmogorov-Smirnov test showed that none of the distributions of the features described in Section 4 were normally distributed. Hence, and similar

Table 3. Means and standard errors of the features in the online social network and the location-based social network for the group of users having reciprocal interactions vs. the groups of users having no reciprocal interactions with each other (*=significant at level 0.1, **=significant at level 0.01, and ***=significant at level 0.001).

	Features	Reciprocal	Non Reciprocal
Online Social Network	Common Neighbors $F_{CN}(u,v)^{***}$	10.20 ± 1.10	0.80 ± 0.10
	Adamic Adar $F_{AA}(u,v)^{***}$	6.46 ± 0.61	0.71 ± 0.06
	Jaccard's Coefficient $F_{JC}(u,v)^{***}$	0.08 ± 0.00	0.04 ± 0.00
	Preferential Attachment $F_{PS}(u,v)^{***}$	12544.28 ± 2066.82	403.15 ± 93.73
	Common Groups $G_C(u,v)$	2.04 ± 0.11	1.81 ± 0.10
	Jaccard's Coefficient $G_{JC}(u,v)$	0.06 ± 0.00	0.05 ± 0.00
	Common Interests $I_C(u,v)$	0.12 ± 0.02	0.05 ± 0.01
	Jaccard's Coefficient $I_{JC}(u,v)$	0.01 ± 0.00	0.00 ± 0.00
Location-based Social Network	Common Neighbors $M_{CN}(u,v)^{***}$	42.59 ± 2.67	61.29 ± 11.96
	Jaccard's Coefficient $M_{JC}(u,v)^{**}$	0.2 ± 0.01	0.19 ± 0.01
	Preferential Attachment $M_{PS}(u,v)^{*}$	41663.58 ± 4547.60	473151.99 ± 411215.48
	Adamic Adar $M_{AA}(u,v)$	21.01 ± 1.30	32.25 ± 7.79
	Regions Seen $R_S(u,v)$	2.82 ± 0.10	2.71 ± 0.18
	Common Regions $R_C(u,v)$	3.25 ± 0.12	4.00 ± 0.83
	Observations Together $R_O(u,v)$	0.23 ± 0.00	0.21 ± 0.00
	Distance $A_D(u,v)^{**}$	9.35 ± 0.48	11.19 ± 0.57
	Days Seen $A_S(u,v)^{**}$	7.22 ± 0.31	6.96 ± 0.33

to Bischoff [3], we compared the variances of all features using a Levene test ($p < 0.01$). To test for significant differences of the means, we employed Mann-Whitney-Wilcoxon test in case of equal variances and a two-sided Kolmogorov-Smirnov test in case of unequal variances. The differences of the means between the groups of users regarding their interaction type can be found in Table 3 and 2. Overall, we found the following:

- *Interactions:* Mean values of topological features in the online social network were significantly higher for user pairs with interactions compared to users without interactions. For homophilic features, a significant difference between user pairs was observed for features based on group affiliation whereas features based on specified interests did not show significant differences. Topological features in the location-based social network also showed significant differences between users but contrary, users with no interactions had a higher number of common neighbors, preferential attachment score, and Adamic Adar score. Users with interactions had more common regions and observations, and they saw each other on more days. Furthermore, user pairs with interactions in the online social network had a significantly shorter average distance between them.
- *Reciprocity:* The differences between user pairs with reciprocal interactions and non-reciprocal interactions can be found in Table 3. The results revealed significant differences between users in the online social network for all topological features but no significant differences for homophilic features. Comparing differences between user pairs also showed significant differences in the

Table 4. Overall results AUC and (ACC) of the Logistic Regression learning approach for predicting interactions between users and their reciprocity in the online social network of Second Life using online social network and location-based social network features.

	Feature Sets		Interaction	Reciprocity
Logistic Regression	Online Social Network	Topological	0.878 (71.8%)	0.676 (64.9%)
		Homophilic	0.640 (63.4%)	0.507 (52.5%)
		Combined	**0.863 (76.8%)**	**0.679 (64.8%)**
	Location-based Social Network	Topological	0.858 (76.7%)	0.530 (51.2%)
		Homophilic	0.885 (80.6%)	0.556 (54.4%)
		Combined	**0.919 (84.8%)**	**0.551 (53.5%)**
	All Features		**0.953 (89.6%)**	**0.709 (65.2%)**

topological features of the location-based social network (Common Neighbors, Jaccard's Coefficient and Preferential Attachment Score) but only the average distance between users and the number of days they saw each other was significantly different for the homophilic features

In the remainder of this section we present the results obtained from the two supervised learning experiments described in Section 5. As learning strategy we used the *Logistic Regression* learning algorithm since it can be easily implemented and interpreted [16].

6.1 Predicting Interactions: Online Social Network vs. Location-based Social Network Features

The results of the first experiment can be found in Table 4 where we present the outcome of the prediction model for two different sources of knowledge and the according feature sets.

The values in the table represent the area under the ROC curve (AUC) and the accuracy of the prediction (ACC) as metrics for the predictability with a baseline for the binary classification problem of 0.5 AUC. As we can see, using topological features from the online social network improved the predictability of interactions between users by +37.8% whereas homophilic features (groups and interests) enhanced the baseline by +14.0%. In contrast to this, topological features from the location-based social network improved the baseline by +35.8% whereas homophilic features improved it by +38.5%. The combined topological and homophilic features from either networks resulted in a predictability of 0.953 AUC outperforming the baseline by +45.3%.

Overall, and interestingly, looking at the feature set in Table 4 we can see that location-based features were a great source to predict interactions between users in online social networks and they even outperformed online social network features. To evaluate the predictability of interactions of features separately, we present the coefficients of the Logistic Regression algorithm and their levels of

Table 5. Coefficients of the Logistic Regression when all topological and homophilic features from both domains are used simultaneously in the dataset (***=significant at level 0.001).

	Features	Interactions	Reciprocity
Online Social Network	Common Neighbors (in) $F^{-}_{CN}(u,v)$	-1.782615***	–
	Common Neighbors (out) $F^{+}_{CN}(u,v)$	0.138448***	–
	Common Neighbors $F_{CN}(u,v)$	–	**-0.658291***
	Adamic Adar $F_{AA}(u,v)$	0.196078	-0.108824***
	Jaccard's Coefficient (in) $F^{-}_{JC}(u,v)$	0.025060***	–
	Jaccard's Coefficient (out) $F^{+}_{JC}(u,v)$	2.416276 ***	–
	Jaccard's Coefficient $F_{JC}(u,v)$	–	0.495911***
	Preferential Attachment (in) $F^{-}_{PS}(u,v)$	**7.405495***	–
	Preferential Attachment (out) $F^{+}_{PS}(u,v)$	-0.000097	–
	Preferential Attachment $F_{PS}(u,v)$	–	-1.107698
	Common Groups $G_C(u,v)$	-0.000066***	-0.000040***
	Jaccard's Coefficient $G_{JC}(u,v)$	0.216582***	-0.046399
	Common Interests $I_C(u,v)$	-1.230746	1.732937
	Jaccard's Coefficient $I_{JC}(u,v)$	0.932973	7.158616
Location-based Social Network	Common Neighbors $M_{CN}(u,v)$	-0.019859***	-0.004276
	Jaccard's Coefficient $M_{JC}(u,v)$	-0.001736***	-0.000470***
	Preferential Attachment $M_{PS}(u,v)$	0.000551***	0.000574
	Adamic Adar $M_{AA}(u,v)$	0.000001***	0.000000
	Regions Seen $R_S(u,v)$	0.294520	-0.101258
	Common Regions $R_C(u,v)$	0.717518***	0.093925
	Observations Together $R_O(u,v)$	0.022711***	-0.064381
	Distance $A_D(u,v)$	**10.570453***	**1.158166***
	Days Seen $A_S(u,v)$	-0.010596***	-0.002153

significance when all features were used simultaneously. Table 5 shows that Preferential Attachment Score for incoming messages $F^{-}_{PS}(u,v)$ in the online social network and the average distance between users $A_D(u,v)$ in the location-based social network were most impacting features. To give an overview of the correlation of the features, we calculated the pair-wise Spearman-rank correlation of the used features from both domains as shown inTable 6.

6.2 Predicting Reciprocity: Online Social Network vs. Location-based Social Network Features

The results of the second experiment can be found in Table 4 where we present the area under the ROC curve (AUC) and the accuracy of the prediction (ACC). As in the previous experiment the baseline for randomly guessing is 0.5 AUC due to the balanced dataset.

Using topological features from the online social network increased the predictability of reciprocity by +17.6% whereas homophilic features alone (groups

Table 6. Spearman's Correlation Matrix (*=significant at level 0.1, **=significant at level 0.01, and ***=significant at level 0.001).

		F^-_{CN}	F^+_{CN}	F_{AA}	F^-_{JC}	F^+_{JC}	F^-_{PS}	F^+_{PS}	G_C	G_{JC}	I_C	I_{JC}	M_{CN}	M_{JC}	M_{PS}	M_{AA}	R_S	R_C	R_O	A_D	A_S
Online Social Network	F^-_{CN}	1.00																			
	F^+_{CN}	0.55***	1.00																		
	F_{AA}	0.49***	0.41***	1.00																	
	F^-_{JC}	0.99***	0.51***	0.47***	1.00																
	F^+_{JC}	0.53***	1.00***	0.40***	0.50***	1.00															
	F^-_{PS}	0.48***	0.47***	0.47***	0.46***	0.46***	1.00														
	F^+_{PS}	0.44***	0.49***	0.56***	0.41***	0.47***	0.30***	1.00													
	G_C	0.16***	0.08***	0.09***	0.17***	0.09***	0.19***	0.06***	1.00												
	G_{JC}	0.16***	0.08***	0.08***	0.17***	0.09***	0.18***	0.06***	0.99***	1.00											
	I_C	0.11***	0.13***	0.12***	0.10***	0.12***	0.13***	0.12***	0.04**	0.03*	1.00										
	I_{JC}	0.11***	0.13***	0.12***	0.10***	0.12***	0.13***	0.12***	0.04*	0.03*	1.00***	1.00									
Location-based Social Network	M_{CN}	0.13***	0.09***	0.08***	0.14***	0.10***	**0.22****	0.06***	**0.20****	**0.21****	0.02	0.02	1.00								
	M_{JC}	0.03*	0.01	0.01	0.04*	0.01	0.05***	0.02	0.11***	0.10***	0.01	0.01	0.74***	1.00							
	M_{PS}	0.18***	0.14***	0.07***	0.19***	0.15***	**0.27****	0.07***	**0.24****	**0.26****	0.03*	0.03*	0.45***	0.32***	1.00						
	M_{AA}	-0.16***	-0.11***	-0.10***	-0.17***	-0.12***	**-0.30****	-0.07***	**-0.19****	**-0.20****	-0.04**	-0.04**	-0.38***	-0.19***	-0.45***	1.00					
	R_S	**-0.20****	-0.16***	-0.14***	**-0.21****	-0.16***	**-0.35****	-0.10***	**-0.22****	**-0.23****	-0.03*	-0.03*	-0.22***	0.10***	-0.53***	0.52***	1.00				
	R_C	-0.16***	-0.13***	-0.08***	-0.17***	-0.14***	**-0.24****	-0.09***	-0.19***	**-0.20****	-0.01	-0.01	-0.22***	-0.05***	-0.51***	0.32***	0.52***	1.00			
	R_O	-0.18***	-0.14***	-0.12***	-0.19***	-0.15***	**-0.32****	-0.08***	-0.18***	-0.19***	-0.03*	-0.03*	-0.18***	0.16***	-0.47***	0.50***	0.94***	0.34***	1.00		
	A_D	-0.07***	-0.05***	-0.06***	-0.07***	-0.06***	-0.17***	-0.02*	-0.02*	-0.03*	-0.01	-0.01	-0.08***	0.14***	-0.16***	0.33***	0.61***	-0.14***	0.78***	1.00	
	A_S	0.14***	0.11***	0.08***	0.14***	0.12***	0.16***	0.08***	**0.20****	**0.20****	0.02	0.02	0.38***	0.31***	0.21***	-0.06***	0.10***	-0.30***	0.23***	0.51***	1.00
		Online Social Network											Location-based Social Network								

and interests) performed as bad as the baseline. Due to the little predictive power of the homophilic features the combination of all features in the online social network results in a prediction gain of +17.6% which is equal to topological features alone. In contrast to this, topological features from the location-based social network improved the baseline approach by +3.0% for the topological features and by +5.6% for the homophilic features. The combination of feature sets in the location-based social network boosted the predictability by +5.1%. The combination of features from either domains elevated the predictability of the reciprocity between two users up to 0.709 AUC, which is a boost of +20.9% if compared to the baseline of 0.5 AUC. Similar to the previous experiment, we computed the coefficients of the Logistic Regression algorithm in Table 5. In the online social network domain the Common Neighbors feature $F_{CN}(u, v)$ and in the location-based social network domain the distance between users $A_D(u, v)$ had the highest and most significant values.

6.3 Verification of Stability: Predicting Interactions and Reciprocity with SVM and Random Forrest

The results of the conducted experiments based on LogisticRegression clearly showed that features from the location-based social network are better suited to predict interactions between users, whereas features from the online social network are better suited to predict reciprocity of interactions. However, to verify the stability of these findings we employed two additional learning algorithms: *Random Forest* and *Support Vector Machine* which are well suited for high dimensional, numeric and inter-dependent attributes (see Table 6) [3], [16]. The results of these learning algorithms are presented in Tables 7 and 8. Overall, the results can be interpreted as follows:

- *Predicting Interactions:* Using Logistic Regression, features from the location-based social network outperformed features from the online social network and similar results were observed for *Support Vector Machine* and *Random Forest*. In both cases features of the location-based social network resulted in a better prediction of interactions than features from the online social network. Overall, the performance of the combined feature set using Support Vector Machine was 0.882 AUC and using Random Forest was 0.979 AUC.
- *Predicting Reciprocity:* For the prediction of reciprocity of interactions between users using Logistic Regression, online social network features outperformed location-based social network features. For other learning algorithms we found similar results as features from the online social network also outperformed features from the location-based social network. The combination of all features from both domains predicted reciprocity of interactions with 0.652 AUC using Support Vector Machine respectively 0.684 using Random Forest.

Table 7. Overall results AUC and (ACC) of the SVM learning approach for predicting interactions between users and their reciprocity in the online social network of Second Life using online social network and location-based social network features.

	Feature Sets		Interactions	Reciprocity
SVM	Online Social Network	Topological	0.669 (66.9%)	0.646 (64.6%)
		Homophilic	0.638 (63.8%)	0.522 (52.2%)
		Combined	**0.737 (73.7%)**	**0.639 (63.9%)**
	Location-based Social Network	Topological	0.793 (79.3%)	0.529 (52.9%)
		Homophilic	0.761 (76.1%)	0.515 (51.5%)
		Combined	**0.849 (84.9%)**	**0.539 (53.9%)**
	All Features		**0.882 (88.2%)**	**0.638 (63.8%)**

7 Discussion and Conclusions

In this work we harvested data from two Second Life related data sources: an online social network with text-based interactions and a location-based social network with position data. We modeled the social proximity between users with topological and homophilic network features and conducted two experiments.

To answer the first research question *RQ1*, we compared different features of user pairs regarding their interactions and the reciprocity of these interactions. This analysis revealed that pairs with interactions were tighter connected in the online social network but the opposite was observed for the location-based social network. A possible explanation is that users in Second Life are allowed to directly "jump" to different regions in the whole virtual world but see the present users only upon arrival. We believe that users are more likely to stay in a region if they know present users, i.e. they have interactions on the online social network. This mobility activity could explain the tight connections in the location-based social network. This assumption is supported by homophilic features from both networks: users with interactions had more common groups, regions, and they saw each other on more days. Furthermore, the average distance was significantly shorter than users without interactions. All observed features were significantly different except interest based features but we assume this is due to the sparse data. The found results for predicting reciprocity of interactions was similar to the prediction of interactions themselves. User pairs with reciprocal interactions had tight connections in the online social network but the opposite was observed for the location-based social network. Again, homophilic features of user pairs with reciprocal interactions indicated a higher alikeness in both networks.

For the second research question *RQ2* we predicted interactions and the reciprocity of these interactions. To do so, we chose Logistic Regression because it is easy to implement and interpret. We observed that interactions can be better predicted with features from the location-based social network than with features from the social network. Surprisingly, the opposite was observed for the reciprocity of interactions. In both experiments we found the combination of features from both networks outperforming either networks: Interactions could

Table 8. Overall results AUC and (ACC) of the Random Forrest learning approach for predicting interactions between users and their reciprocity in the online social network of Second Life using online social network and location-based social network features.

	Feature Sets		Interactions	Reciprocity
Random Forest	Online Social Network	Topological	0.893 (79.7%)	0.628 (62.2%)
		Homophilic	0.624 (62.8%)	0.488 (50.4%)
		Combined	**0.910 (82.5%)**	**0.635 (60.5%)**
	Location-based Social Network	Topological	0.852 (77.9%)	0.530 (52.2%)
		Homophilic	0.872 (80.3%)	0.479 (49.2%)
		Combined	**0.916 (85.7%)**	**0.550 (53.2%)**
	All Features		**0.979 (93.0%)**	**0.684 (62.8%)**

be predicted with 0.953 AUC and the reciprocity of these interactions with 0.709 AUC. The Logistic Regression coefficients of the features unveiled that a short average distance between users is a good indicator for interactions and their reciprocity. To verify our results that online social network features outperform features from the location-based social network for the prediction of interactions and vice versa for the prediction of reciprocity, we used two additional learning algorithms: Support Vector Machines and the Random Forest learning approach. Both algorithms approved the observations made in the experiment with Logistic Regression.

To answer the third research question *RQ3*, we compared homophilic features and topological features regarding the predictability of interactions and their reciprocity. Interestingly, we could not find a stable pattern over all experiments, as it was for instance proposed by Rowe *et al.* [19]. Although topological features of the online social network outperformed homophilic features in all three learning algorithms we found variation of the results for the location-based social network. Using Logistic Regression homophilic features performed better than topological features but in contrast, the opposite was observed for Support Vector Machines. With Random Forest homophilic features were better suited for the prediction of interactions but homophilic features were better suited for the reciprocity of interactions.

For future work, it is planned to dig deeper into the data and to address issues such as the variety of time (which we did not address in this study) or the issue why reciprocal links seem to be better predicted with social network features than with position data. Furthermore, we plan to extend our approach to predict other relations between users besides communicational interactions such as for instance partnership which can be also mined from the social network of Second Life. Finally, it is our interest to switch from supervised to unsupervised learning.

Acknowledgements

This work is supported by the Know-Center. The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency (FFG).

References

1. L. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.
2. A. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.
3. K. Bischoff. We love rock'n'roll: analyzing and predicting friendship links in Last. fm. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 47–56. ACM, 2012.
4. J. Cheng, D. Romero, B. Meeder, and J. Kleinberg. Predicting reciprocity in social networks. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 49–56. IEEE, 2011.
5. A. Clauset, C. R. Shalizi, and M. Newman. Power-law distributions in empirical data, 2007. *arXiv preprint arXiv:0706.1062*, 64.
6. J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 119–128. ACM, 2010.
7. M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici. Link prediction in social networks using computationally efficient topological features. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 73–80. IEEE, 2011.
8. S. Golder and S. Yardi. Structural predictors of tie formation in twitter: Transitivity and mutuality. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 88–95. IEEE, 2010.
9. R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web*, pages 403–412. ACM, 2004.
10. I. Guy, M. Jacovi, A. Perer, I. Ronen, and E. Uziel. Same places, same things, same people?: mining user similarity on social media. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, pages 41–50, New York, NY, USA, 2010. ACM.
11. I. Guy, M. Jacovi, E. Shahar, N. Meshulam, V. Soroka, and S. Farrell. Harvesting with sonar: the value of aggregating social network information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1017–1026, New York, NY, USA, 2008. ACM.
12. I. Guy, N. Zwerdling, D. Carmel, I. Ronen, E. Uziel, S. Yogev, and S. Ofek-Koifman. Personalized recommendation of social software items based on social relations. In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, pages 53–60, New York, NY, USA, 2009. ACM.

13. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.
14. J. Huang and C. X. Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Trans. on Knowl. and Data Eng.*, 17(3):299–310, Mar. 2005.
15. A. Jain and R. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.
16. J. J. Jones, J. E. Settle, R. M. Bond, C. J. Fariss, C. Marlow, and J. H. Fowler. Inferring tie strength from online directed behavior. *PloS one*, 8(1):e52168, 2013.
17. D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.
18. C. X. Ling, J. Huang, and H. Zhang. Auc: a statistically consistent and more discriminating measure than accuracy. In *International Joint Conference on Artificial Intelligence*, volume 18, pages 519–526. LAWRENCE ERLBAUM ASSOCIATES LTD, 2003.
19. M. Rowe, M. Stankovic, and H. Alani. Who will follow whom? exploiting semantics for link prediction in attention-information networks. 2012.
20. S. Scellato, A. Noulas, and C. Mascolo. 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*, pages 1046–1054. ACM, 2011.
21. M. Steurer, C. Trattner, and F. Kappe. Success factors of events in virtual worlds a case study in second life. In *NetGames*, pages 1–2, 2012.
22. D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. ACM, 2011.
23. D. Yin, L. Hong, and B. Davison. Structural link analysis and prediction in microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1163–1168. ACM, 2011.