# Predicting Interactions In Online Social Networks: An Experiment in Second Life

Michael Steurer
Institute for Information Systems
and Computer Media
Graz University of Technology
Graz, Austria
msteurer@iicm.tugraz.at

Christoph Trattner
Know-Center
Graz University of Technology
Graz, Austria
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, studies exploiting different kinds of knowledge sources for the link prediction problem are rare. In this paper latest results of a project are presented 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 social network and position data. To that end, we conducted two experiments in the virtual world of Second Life. As our results reveal, position data features are a great source to predict interacts between users in online social networks and outperform social network features significantly. However, if we try to predict reciprocal message communication between users, social network features seem to be superior.

## Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*

## General Terms

Measurement, Experimentation

## Keywords

online social network, location-based, link prediction, virtual worlds


*MSM'13*, May 1, 2013, Paris, France.


## 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 [14], *e.g.* the issue of forecasting whether or not two users $u$ and $v$ of a given social network $G\langle V, E\rangle$, will interact with each other in the future. While considerable amount of work has been conducted recently of how to predict links between users in online social media, comparing different sources of knowledge with each other 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 using both social network and position data. To tackle this issue we train a binary classifier that learns the relations between users $u$ and $v$ based on a number of features induced from social network and position data. For the purpose of our study we furthermore differentiate between two types of features – network topological and homophilic features [18]. Since it is nearly impossible to obtain rich large-scale real-world social network and position data, our investigation focuses on the virtual world of Second Life, where we can 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 can interact with each other by sharing text messages, comments, and loves. Additionally, we were able to collect position data of residents in the virtual world by implementing so-called in-world bots. These bots collect tracking information with accurate timestamp and position information of surrounding residents and store it persistently. Overall, it is our interest to answer the following research questions in this paper:

- To what extent can we predict interactions between users in online social networks using social network or position data?
- To what extent can we enhance the predictability of interactions between users in online social networks combining both – social network and position data?

To that end, we conducted two experiments in the virtual

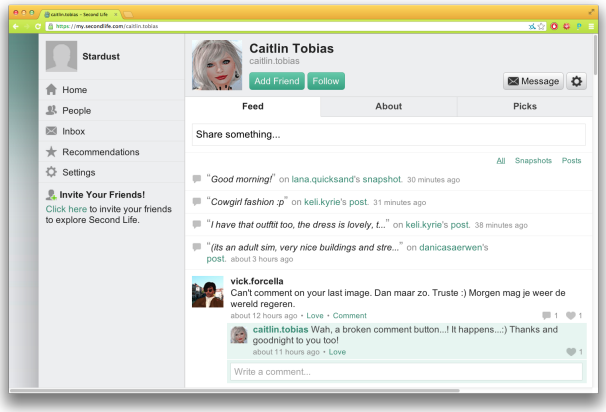


**Figure 1: Sample of a user profile in the social network *My Second Life*. As shown, users can *post* text message on their wall or can communicate with each other by *commenting* or *loving* onto each other's posts.**

world of Second Life. As our results reveal, position data features are a great source to predict interacts between users in online social networks outperforming social network features to a great extent. However, if we try to predict reciprocal message communication between users, social network features seem to be superior.

The remainder of 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 Section 8 and Section 9 concludes the paper.

## 2. RELATED WORK

Although considerable amount of work has been conducted recently 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 GPS data and Facebook friendship data through a mobile app [4]. Based on a number of experiments they show that the so-called place-entropy feature is 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.* [9, 10, 8] 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 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 [6]. In their paper they study the microblogging 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.* [3]. By applying a rich set of network based features including link prediction features from [14], 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* [19]. 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.* [5]. In their work they study a rich set of features (over 20) 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 [15] they study the link prediction problem, or the question who will follow whom, in the microblogging system *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 Weibo. Hence, contrary to our own work they did not include external knowledge such as position data as we do in our study. In contrast to the work of Rowe *et al.*, the last study to be mentioned is by Scellato *et al.* [16] who tried to exploit features from the location-based social network of *Gowalla* to predict links. In contrast to our work, they only focused on position data of users and did not combine the data with an additional network of interactions 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 – social network and position 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 the users profiles. Second, contrary to real world online social networks most of the profiles in *My Second Life* are public, *i.e.* we can mine a large fraction of the network. Third, in virtual worlds the position 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 Position Dataset

The collection of position data of users in Second Life is a two stage process: First, a list of popular locations from the Second Life Event calendar[1] were crawled. Second, overall 15 in-world agents, the so-called in-world-bots, were implemented to teleport to these locations to gather position information about the present users.

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. Overall, we were able to obtain data of 218,245 unique events during a period of ten months starting in March 2012.

To collect position data of the users we implemented 15 in-world agents on the basis of the open source command-line client *libopenmetaverse*[2]. 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 position data of all surrounding users in a region. This position information comprises the current region, x and y coordinates of the position within this region, and a timestamp (see Figure 2). 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 is 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 [17].

## 3.2 Social Network Dataset

In July 2011 Linden Labs introduced an online social network called *My Second Life*[3] similar to other online 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.

Overall, we downloaded the profile data of 190,160 users found by our in-world agents, parsed the interaction partners of the these users and downloaded the profile information of the missing ones. This procedure was repeated until the crawler could not find any more new users. Finally, this approach yields in a dataset of 311,959 users with 135,181 having interactions on their feed.

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

**Figure 2: A map section in Second Life with users represented as white dots. The in-world agent visiting the region periodically is outlined as a crossed white dot.**

# 4. FEATURE SETS

As already outlined, it is our aim to predict interactions between users in online social networks based on two types of knowledge sources – social network and position data of the users. To that end, we induce two different types of feature sets from our data sources: network topological and homophilic features [18]. 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 *social network*, is based on the data obtained from the users profile where every edge in this directed network indicates communication between two users. This yields in a network with 135,181 users and 209,653 edges. The second network, referred to as *position network*, is based on the users position data where every edge in the undirected network indicates that two users were seen concurrently in the same region on two different days. This yields in a network with 131,349 nodes and 2,343,683 edges. A summary of both networks can be found in Table 1.

## 4.1 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 consider the social network as a directed graph $G_F\langle V_F, E_F\rangle$ with $V_F$ representing the users where $e = (u, v) \in E_F$ if user $u$ posted, commented, or liked something on the feed of user $v$.

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

- *Common Neighbors* $F_{CN}(u, v)$. This is the number of

**Table 1: Networks used for the experiments**

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

interaction-partners two users have in common.

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

For a directed network we can split this into the number of common users $F^+_{CN}(u,v) = |\Gamma^+(u) \cap \Gamma^+(v)|$ to which both users send messages to and the number of users $F^-_{CN}(u,v) = |\Gamma^-(u) \cap \Gamma^-(v)|$ from which both users receive 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 is taken from [12] and is defined as follows.

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

For directed networks this can be split into two coefficients for receiving messages $F^-_{JC}(u,v) = \frac{|\Gamma^-(u) \cap \Gamma^-(v)|}{|\Gamma^-(u) \cup \Gamma^-(v)|}$ and sending 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 [3].

- *Preferential Attachment Score* $F_{PS}(u,v)$. This feature takes 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 can be applied to a directed network with two different features: $F^+_{PS}(u,v) = |\Gamma^+(u)| \cdot |\Gamma^-(v)|$, respectively $F^-_{PS}(u,v) = |\Gamma^-(u)| \cdot |\Gamma^+(v)|$ [3].

## 4.2 Social Network: Homophilic Features

As stated before, users in Second Life can enhance their 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 [15, 18], homophily is an important variable in the context of the link prediction problem. To account for factor, we define a set of homophilic features which we calculate as group and interest similarity between users $u, v$. Formally, we define 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 is 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, is the number of interests two users have 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 Position Network: Topological Features

We can apply the same network topological feature calculations to the position network as we did for the social network. The network has edges between users that met on at least two days. Using this relation between in-world users we can define the topological features similar to Section 4.1. Here, the neighbors of a node in the undirected position network $G_M\langle V_M, E_M\rangle$ are defined as $\Theta(u) = \{v \mid (u,v) \in G_M\}$ and starting with this we define 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 Position Network: Homophilic Features

These features are 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 reside in the same region concurrently. An observation $o \in O(u, v)$ is 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* $M_D(u, v)$. Whenever two users are observed concurrently, we can measure the distance between them based on their x and y coordinates. As a measure for their overall physical closeness, we can therefore compute the average physical Euclidean distance between two users for all observations where both are present in the same region concurrently.

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

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

The regions of a user are defined as $P(u) = \{\rho \in P|$ user $u$ was observed in region $P\}$ and so we can compute 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 takes only the regions into account where both users have been observed in the same region concurrently.

- *Observations Together* $R_O(u, v)$. This feature is taken from Cranshaw *et al.* [4] and represents 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

We conducted two different experiments using two different datasets and features to study the extent to which interactions between users in online social networks can be predicted. These experiments are based on the *social network* $G_F$ and the *position network* $G_M$ dataset as described in Section 4.

In the first experiment we try to predict the interactions between users in the social network and in the second experiment we try to predict whether these links are reciprocal or not. In both experiments we follow the approach of Guha *et al.* [7] who suggest to create two datasets with an equal number of "positive edges" and "negative edges" for the binary classification problem which yields 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 measures we follow the approaches of Leskovec and Rowe *et al.* [13, 15], *i.e.* we employed the binomial *Logistic Regression* algorithm and use the area under the ROC curve (AUC) and the accuracy of the prediction (ACC) as evaluation metrics. Furthermore, we used a 10-fold cross validation approach to justify our findings. For both experiments we used the binomial Logistic Regression algorithm of the WEKA machine-learning suite [11]. In the following sections we describe in detail how the trainings- and test data for both experiments were generated.

## 5.1 Predict Interactions

The task here is to predict whether two users interact with each other on the feed by exploiting topological and homophilic information of the social network and the position network. The experiment is based on the combination of the directed social network $G_F\langle V_F, E_F\rangle$ and the undirected positions network $G_M\langle V_M, E_M\rangle$. In the first step we compute the edge-features for the user-pairs as described in Section 4 for both networks independently. Then, in the second step we create the intersection of both networks as a directed graph $G_{FM}\langle V_{FM}, E_{FM}\rangle$ where $V_{FM} = \{v|v \in V_F, v \in V_M\}$, and $E_{FM} = \{(u, v)|(u, v) \in E_M, (u, v) \in E_F, v$ and $u \in V_{FM}\}$. This newly created network consists of 37,118 nodes and 1,014,352 pairs with position 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 select 2,500 user-pairs with a social interaction and a position 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 position 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 are used as data set for the learning algorithm.

## 5.2 Predict Reciprocity

The task here is to predict whether two users have mutual activities on each other's wall, *i.e.* reciprocal interactions, by exploiting topological and homophilic information of the social network and the position network. We define 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 use this difference for the binary classification problem. In contrast to the previous experiment we consider the social network as undirected network for the computation of the edge-features described in Section 4 but retain information about the reciprocity of the interactions. The edge features for the position network are again considered as undirected. For the actual experiment we combine the social network and the position 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 $E'_{FM} = \{(u, v)|(u, v) \in E_M, (u, v) \in E_F$ or $(v, u) \in E_F, v$ and $u \in V'_{FM}\}$. Out of the 7,393 user-pairs with a social interaction and a position co-occurrence we could identify 1,431 reciprocal edges and 4,531 non-reciprocal edges in the 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

**Table 2: Overall results of the area under the ROC curve (AUC) and the accuracy (ACC) for predicting interactions and reciprocity between users in the online social network of Second Life using social network and position network features.**

| Feature Sets | | Interactions | Reciprocity |
|---|---|---|---|
| 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%)** |
| Position 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%)** |

**Table 3: Detailed results of the area under the ROC curve (AUC) and the accuracy (ACC) for predicting interactions and reciprocity between users in the online social network of Second Life using only social network features.**

| | Features | | Predict Interaction | Predict Reciprocity |
|---|---|---|---|---|
| *Social Network* | Topological Features | Common Neighbors $F_{CN}(u,v)$ | - | 0.678 (64.4%) |
| | | Common Neighbors (out) $F^{+}_{CN}(u,v)$ | 0.606 (61.0%) | - |
| | | Common Neighbors (in) $F^{-}_{CN}(u,v)$ | 0.626 (63.1%) | - |
| | | Adamic Adar $F_{AA}(u,v)$ | - | **0.681 (65.4%)** |
| | | Adamic Adar (in) $F^{-}_{AA}(u,v)$ | 0.656 (62.8%) | - |
| | | Jaccard's Coefficient $F_{JC}(u,v)$ | - | 0.651 (61.6%) |
| | | Jaccard's Coefficient (out) $F^{+}_{JC}(u,v)$ | **0.888 (70.5%)** | - |
| | | Jaccard's Coefficient (in) $F^{-}_{JC}(u,v)$ | 0.606 (60.7%) | - |
| | | Preferential Attachment $F_{PS}(u,v)$ | - | 0.630 (55.0%) |
| | | Preferential Attachment (out) $F^{+}_{PS}(u,v)$ | **0.888 (70.5%)** | - |
| | | Preferential Attachment (in) $F^{-}_{PS}(u,v)$ | 0.619 (60.7%) | - |
| | Homophilic Features | Common Groups $G_C(u,v)$ | 0.629 (60.8%) | 0.494 (50.8%) |
| | | Jaccard's Coefficient $G_{JC}(u,v)$ | **0.633 (62.8%)** | 0.503 (51.4%) |
| | | Common Interests $I_C(u,v)$ | 0.510 (51.3%) | **0.511 (52.0%)** |
| | | Jaccard's Coefficient $I_{JC}(u,v)$ | 0.511 (51.4%) | **0.511 (52.0%)** |

("negative edges") each. The edges, *i.e.* user-pairs with the according features, are again used for the learning algorithm.

# 6. RESULTS

In this section we present the results obtained from the two experiments.

## 6.1 Predict Interactions: Social Network vs. Position Network Features

The results of the first experiment can be found in Table 2 where we show the differences and similarities between the two sources of knowledge and the features. 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. The baseline for the binary classification problem is 0.5 (AUC). As we can see, using topological features for our classifier from the social network improves the predictability of interactions between users by +37.8% whereas homophilic features (groups and interests) enhance the baseline by +14.0%. In contrast to this, we can see that topological features from the position network improve the baseline approach by +35.8% whereas homophilic features improve the baseline by +38.5%.

Overall, and interestingly, looking at the "combined feature" set in Table 2 we can see that position data features are a great source to predict interactions between users in online social networks outperforming social network features significantly. In Table 3 and 4 one can find an overview of each single feature, with best the features highlighted in bold letters.

## 6.2 Predict Reciprocity: Social Network vs. Position Network Features

The results of the second experiment can be found in Table 2 where we again present the area under the ROC curve (AUC) and the accuracy of the prediction (ACC). Due to the balanced dataset our baseline is again 0.5 (AUC).

As we can see, using topological features for our classifier improves the predictability of interactions between users by +17.6% whereas homophilic features alone (groups and interests) perform as bad as the baseline approach. In contrast to this, we can see that topological features from the position network improve the baseline approach by +3.0% for the topological features and by +5.6% for the homophilic features.

Interestingly, and contrary to the predictability of interactions, we can see from the "combined feature" set in Table 2 that social network features seem to be superior to position network features, if we try to predict the reciprocity of users. In Table 3 and 4 we also present the predictive power of each single feature used for our experiments. Best

**Table 4: Detailed results of the area under the ROC curve (AUC) and the accuracy (ACC) for predicting interactions and reciprocity between users in the online social network of Second Life using only position network features.**

| Features | | | Predict Interaction | Predict Reciprocity |
|---|---|---|---|---|
| *Position Network* | Topological Features | Common Neighbors $M_{CN}(u,v)$ | 0.566 (51.5%) | **0.539 (51.7%)** |
| | | Jaccard's Coefficient $M_{JC}(u,v)$ | 0.734 (59.1%) | 0.521 (51.2%) |
| | | Preferential Attachment $M_{PS}(u,v)$ | **0.754 (60.5%)** | 0.520 (51.7%) |
| | | Adamic Adar $M_{AA}(u,v)$ | 0.662 (58.1%) | 0.508 (51.6%) |
| | Homophilic Features | Regions Seen $R_S(u,v)$ | 0.708 (69.5%) | 0.513 (51.1%) |
| | | Common Regions $R_C(u,v)$ | 0.566 (51.5%) | 0.503 (50.5%) |
| | | Observations Together $R_O(u,v)$ | 0.800 (72.2%) | **0.543 (53.3%)** |
| | | Distance $M_D(u,v)$ | **0.821 (66.5%)** | 0.534 (52.8%) |
| | | Days Seen $M_D(u,v)$ | 0.627 (59.0%) | 0.496 (50.8%) |

features are again highlighted with bold letters.

# 7. DISCUSSIONS

In conclusion, the results of both experiments show that the predictability of interactions and reciprocity between users in the social network of Second Life can be significantly improved if we train our classifier on both sets of features – social network and position network features. Furthermore, we observe that interactions can be better predicted 0.953 (AUC) than reciprocity 0.709 (AUC). What is also interesting to note is the fact that social network topological features perform better than social network homophilic features for predicting interactions and reciprocity. The opposite could be observed for position network topological and homophilic features.

Besides the Logistic Regression approach suggested in [13, 15] we also tested other learning strategies such as Support Vector Machines and Decision Trees (C4.5) as described in [5] for both experiments. However, for the sake of space we present only the best results which we obtained with the Logistic Regression approach.

# 8. CONCLUSIONS

In this paper we presented latest results of a project that studies the extent to which interactions between users in online social networks can be predicted exploring features obtained from social network and position data. To that end, we conducted two experiments in the virtual world of Second Life. As our results revealed, position data features are a great source to predict interactions between users in online social networks outperforming social network features significantly. However, if we try to predict reciprocal message communication between users, social network features seem to be superior.

# 9. FUTURE WORK

Overall, we believe that the findings presented in this paper open new perspectives for further research in the scope of virtual worlds as well as in the real world. 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.

# 10. 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).

# 11. 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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] A. Jain and R. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[13] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, pages 641–650. ACM, 2010.

[14] 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.

[15] M. Rowe, M. Stankovic, and H. Alani. Who will follow whom? exploiting semantics for link prediction in attention-information networks. 2012.

[16] 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.

[17] 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.

[18] 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.

[19] 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.