Predicting Social Interactions from Different Sources of Location-based Knowledge

Michael Steurer
Institute for Information Systems and Computer Media
Graz University of Technology
Email: michael.steurer@iicm.tugraz.at

Christoph Trattner
Know-Center
Graz University of Technology
Email: ctrattner@know-center.at

Denis Helic
Knowledge Technology Institute
Graz University of Technology
Email: dhelic@tugraz.at

***Abstract*—Recent research has shown that digital online geo-location traces are new and valuable sources to predict social interactions between users, *e.g.* check-ins via FourSquare or geo-location information in Flickr images. Interestingly, if we look at related work in this area, research studying the extent to which social interactions can be predicted between users by taking more than one location-based knowledge source into account does not exist. To contribute to this field of research, we have collected social interaction data of users in an online social network called My Second Life and three related location-based knowledge sources of these users (monitored locations, shared locations and favored locations), to show the extent to which social interactions between users can be predicted. Using supervised and unsupervised machine learning techniques, we find that on the one hand the same location-based features (*e.g.* the common regions and common observations) perform well across the three different sources. On the other hand, we find that the shared location information is better suited to predict social interactions between users than monitored or favored location information of the user.**

***Keywords*—Social Interaction Prediction; Location-Based Social Networks; Link Prediction; Virtual Worlds**

I. Introduction

There is no doubt that tomorrow's world will be mobile and social. It is therefore not surprising that recent research has rigorously followed this trend to study new methods to predict social ties or links between people in such an environment. Interestingly, if we look at related work in this area (*e.g.* [1], [2], [3]), research studying the extent to which social links can be predicted between users typically takes just one knowledge source into account, *e.g.* online social network data from Facebook, or location-based social network data from FourSquare. To contribute to this emergent and still sparse field of research, we have recently started a project (see [4], [5], [6]) with the overall goal to predict links and tie strength between users from various sources of social and mobile data. Since it is nearly impossible to obtain a complete dataset containing both kinds of knowledge sources in the real world, we focused our experiments on a virtual environment called My Second Life. This allowed us to easily mine any kind of information needed for such a type of a project on a large scale. So far, we have studied the extent to which partnership [6] and in general interactions can be predicted [5] by looking at homophilic features such as for instance common interests, common groups, or common-places visited and network topological features where we investigated common friends features such as Adamic Adar, Jaccard's coefficient etc. Interestingly, we find that the location information of users is to a great extent useful to predict tie strength for the interactions between them in the virtual world of Second Life, most of the time outperforming online social network features. While we only used one particular type of location-based knowledge source about users, namely monitored locations, in our previous research, in this paper we are interested to overall study three different types of knowledge sources: monitored locations, shared locations and favored locations. We employed 10 different features to predict social interactions between users and unveil what type of location-based knowledge source and what types of features were most valuable. Overall, we would like to answer the following research questions in this paper:

RQ1. Are there any statistically significant differences between the users having and not having social interaction with each other based on the features induced from our three different kinds of location-based knowledge sources?

RQ2. Which features perform best across those three types of location-based knowledge sources?

RQ3. What kind of knowledge sources is in the end the most valuable to predict social interactions between users?

To answer the first question we analyzed the datasets with statistical methods according to our features. This evaluation showed that there were significant differences between user-pairs with a social interaction and users without an social interaction across all computed features and all three sources of location information. For instance, user-pairs with a social interaction share more common regions compared to user-pairs without social interaction. To answer the second research question, we employed Collaborative Filtering for each feature independently to predict the social interactions between the users to find the most valuable features. Among others we found that common regions and common observations of two users were a good indicator for an social interaction between them. For the last question we combined the best features for each region source and showed that the user's Shared Locations were more valuable to predict social interactions than Monitored or Favored locations.

In detail the paper is structured as follows: In Section II we shortly discuss related work in the area. In Section III we introduce the collected datasets and the features to predict social interactions between users in Section IV. The setup of the experiments is depicted in Section V followed by the results in Section VI. Finally, Section VII discusses the findings and concludes the paper.

II. Related Work

Approaches by Liben *et al.* [7] or Hasan *et al.* [8] for link prediction using features obtained from online social networks where greatly enhanced with the advent of user's location data. On of the first studies in this field was conducted by Cranshaw *et al.* [2] who combined the interaction of the online social network *Facebook* with the location-based social network of *Loccaccino*. They introduced various metrics to compute users homophily and found a significant correlation between social interactions and location-based features. Similar observations were made by Thelwall *et al.* [9] who revealed significant homophily between interacting users in *MySpace* and even inferred an real-life friendship from the online social network. This goes inline with Bischoff *et al.* [3] who found relations between connections in *Last.FM* and visited music concerts based on demographic, structural and taste-related attributes. Scellato *et al.* [1] investigated in the location-based social network of *Gowalla* and found 30% of newly created links as "place friends". Research by Wang *et al.* [10] follows this direction. They investigated in the prediction of social relations using mobility data obtained from mobile phones and found mobile information significantly outperforming simple network measures. Another paper by Scellato *et al.* [11] focuses on the structural differences between the three location-based social networks of *Brighknight*, *Foursquare*, and *Gowalla*. In contrast to our work, they did not have different location sources for one single online social network and their focus was on the actual spatial distance between user.

III. Data Sets

Our experiments were based on a social interaction dataset of users in an online social network and three independent location-based knowledge sources: *Monitored Locations*, *Shared Locations*, and *Favored Locations* from a virtual world. In particular, we focused in our experiments on a virtual environment called Second Life, where we could easily mine the necessary information needed for the experiments on a large scale (see [4], [5], [6] for more details).

A. Social Interaction Dataset

The online social network *My Second Life* was introduced by Linden Labs, the company behind Second Life, in July 2011. It is a social network that can be compared to Facebook regarding postings and check-ins but aims only at residents of the virtual world: just as in Facebook, residents can interact with each other sharing text messages, and comment or love (similar to a "like" in Facebook) these messages. Figure 1 depicts a typical profile of a user with postings, comments, and loves from others. A user's profile can be accessed with a unique URI derived from the user name and we attempted to download the profile data of over 400,000 users with a web-crawler. We extracted their interaction partners and downloaded the missing profiles iteratively. With this approach, we found 152,509 profiles with interactions on their profile and identified 1,084,002 postings, 459,734 comments and 1,631,568 loves.

B. Location-based Dataset

To predict the social interactions between users we employed location information obtained from three different sources of data.

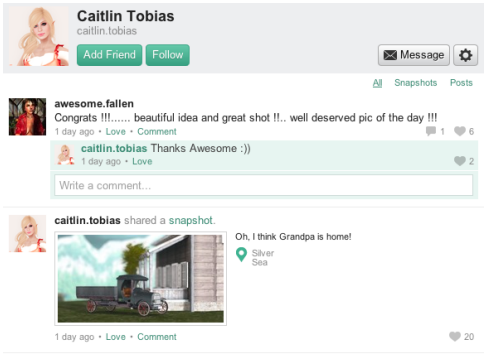

Fig. 1. User profile of an Second Life resident in the online social network My Second Life showing a posting, a shared snapshot with location information (Silver Sea), and a comment.

a) Monitored Locations: As in the real life, residents of Second Life can host events in the virtual world for other residents and publicly announce this information on an event calendar. We implemented a web-crawler that harvested this calendar periodically to extract all events with accurate event-location and start time. Based on this information we have implemented 15 avatar-bots that visited these events with an interval of 15 minutes and collected the accurate location of the participating users. Starting in March 2012 we were able to collect 262,234 events over a period of 12 months yielding in a dataset of nearly 19 million data samples, i.e. user-location tuples, of over 410,616 different users in 4,132 unique locations.

b) Shared Locations: Users of *My Second Life* can not only interact with each other using postings, comments, or loves, they can also share location information about their current in-world location through in-world pictures. The idea of sharing these locations can be compared to pictures uploaded to *Flickr* or *Facebook* enriched with GPS information (see Figure 1). Overall, we identified 496,912 snapshots in 13,583 unique locations on 45,835 profiles.

c) Favored Locations: Every user of Second Life can specify up to 10 so-called "Picks" on it's profile representing the favorite locations of users. Users can enhance these picks with a picture and personal text note. These favored locations are visible to other users and hence it can be easily accessed with a Web browser using a URI derived from the user's name. We found 191,610 profiles, sharing 811,386 locations in 25,311 unique regions.

Figure 2 depicts the number of observations of the collected users for the three location sources. Both, Shared and Monitored Locations show power law qualities which is in contrast to the Favored Locations due to Linden's limitations of 10 picks per user.

IV. Features

Based on the collected location-based user data we induced overall 10 different features in order to measure the homophily between the users and to predict social interactions between them [2], [5], [6]. For the reminder of this paper the sequence of observations $O(u)$ of a user u are denoted as 1) $O_m(u)$ for

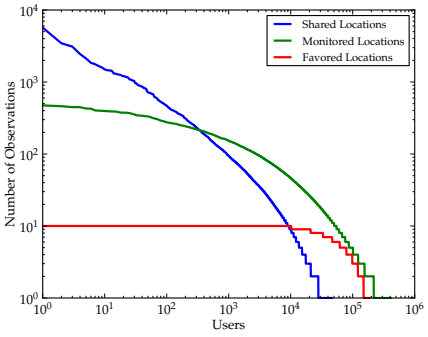

Fig. 2. The number of user observations in the three different location-based knowledge sources.

TABLE I. MEANS AND STANDARD ERRORS OF FEATURES APPLIED TO THE THREE SOURCES OF LOCATION DATA COMPARING USER-PAIRS WITH AND WITHOUT INTERACTIONS ($^{*}p < 0.1$, $^{**}p < 0.01$, AND $^{***}p < 0.001$).

	Features	Have Interactions	Have No Interactions
Monitored Locations	$R_C(u,v)^{***}$	0.49 ± 0.01	0.12 ± 0.00
	$R_{U,\mu}(u,v)^{***}$	179.02 ± 1.49	211.51 ± 2.55
	$R_{U,\sigma}(u,v)^{***}$	188.64 ± 1.17	215.04 ± 1.48
	$R_{E,\mu}(u,v)^{***}$	1.52 ± 0.00	1.60 ± 0.01
	$R_{E,\sigma}(u,v)^{***}$	0.53 ± 0.00	0.56 ± 0.00
	$R_{F,\mu}(u,v)^{***}$	637.50 ± 6.22	755.68 ± 10.96
	$R_{F,\sigma}(u,v)^{***}$	787.40 ± 5.66	894.85 ± 7.71
	$R_{JC}(u,v)^{***}$	0.05 ± 0.00	0.01 ± 0.00
	$R_O(u,v)^{***}$	0.04 ± 0.00	0.01 ± 0.00
	$R_T(u,v)^{***}$	12.24 ± 0.10	10.03 ± 0.07
Shared Locations	$R_C(u,v)^{***}$	1.01 ± 0.02	0.02 ± 0.00
	$R_{U,\mu}(u,v)^{***}$	22.78 ± 0.23	38.10 ± 3.23
	$R_{U,\sigma}(u,v)^{***}$	28.91 ± 0.25	37.09 ± 1.52
	$R_{E,\mu}(u,v)^{**}$	0.80 ± 0.00	0.86 ± 0.02
	$R_{E,\sigma}(u,v)^{***}$	0.44 ± 0.00	0.46 ± 0.01
	$R_{F,\mu}(u,v)^{***}$	92.99 ± 0.82	144.85 ± 11.18
	$R_{F,\sigma}(u,v)^{*}$	160.61 ± 1.14	180.70 ± 7.31
	$R_{JC}(u,v)^{***}$	0.03 ± 0.00	0.00 ± 0.00
	$R_O(u,v)^{***}$	0.02 ± 0.00	0.00 ± 0.00
	$R_T(u,v)^{***}$	63.70 ± 0.59	15.11 ± 0.19
Favored Locations	$R_C(u,v)^{***}$	0.11 ± 0.00	0.00 ± 0.00
	$R_{U,\mu}(u,v)^{***}$	12.90 ± 0.33	18.57 ± 1.76
	$R_{U,\sigma}(u,v)^{***}$	13.23 ± 0.37	22.26 ± 2.17
	$R_{E,\mu}(u,v)^{**}$	0.71 ± 0.01	0.81 ± 0.03
	$R_{E,\sigma}(u,v)^{***}$	0.40 ± 0.00	0.51 ± 0.02
	$R_{F,\mu}(u,v)^{**}$	16.17 ± 0.37	21.55 ± 1.92
	$R_{F,\sigma}(u,v)^{**}$	15.79 ± 0.40	25.01 ± 2.28
	$R_{JC}(u,v)^{***}$	0.02 ± 0.00	0.00 ± 0.00
	$R_O(u,v)^{***}$	0.02 ± 0.00	0.00 ± 0.00
	$R_T(u,v)^{***}$	8.04 ± 0.03	6.95 ± 0.03

Monitored Locations, 2) $O_s(u)$ for Shared Location, and 3) $O_f(u)$ for Favored Locations. In contrast, the set of locations where a user was observed is defined as $P(u) = \{\rho \in O(u)\}$. The actual features we used in our experiments are as follows:

a) Common Locations $R_C(u,v)$*:* The simplest metric to determine the homophiliy between two users u and v is the number of regions they have visited in common. In particular this can be computed as $R_C(u,v) = |P(u) \cap P(v)|$.

b) Total Locations $R_T(u,v)$*:* Analogous to the common regions, one can also define the regions two users have in total and use it as a homophilic feature $R_T(u,v) = |P(u) \cup P(v)|$.

c) Jaccard's Coefficient $R_{JC}(u,v)$*:* A combination of the common regions of two users and their total regions is the overlap of locations which is defined as the fraction of common locations and locations visited by both users [2]. This feature is also known as Jaccard's Coefficient $R_{JC}(u,v) = \frac{|P(u) \cap P(v)|}{|P(u) \cup P(v)|}$.

d) Location Observations $R_O(u,v)$*:* Another feature taken from Cranshaw [2] is the location observations that is similar to the Jaccard's Coefficient between two users. It is computed as the number of locations two users have in common divided by the sum of locations either user have $R_O(u,v) = \frac{|P(u) \cap P(v)|}{|P(u)| + |P(v)|}$.

e) Location User-Count $R_U(u,v)$*:* The following three features were first introduced by Cranshaw *et al.* [2] and try model the location diversity of regions two users visited in common. The first and most simple feature to include the popularity of a region is the overall number of observations of unique users at a certain region. According to this we calculated the mean user-count $R_{U,\mu}(u,v)$ and the standard deviation of the mean $R_{U,\sigma}(u,v)$ of all regions two users visited in common $P(u) \cap P(v)$.

f) Location Frequency $R_F(u,v)$*:* The second feature taken from [2] is similar to the previous feature of counting users at a certain location. We computed the frequency defined as the overall observations of users at a certain location. Again we calculated the mean frequency $R_{F,\mu}(u,v)$ and the according standard deviation $R_{F,\sigma}(u,v)$ of the frequency of regions two users u and v have in common.

g) Location Entropy $R_E(u,v)$*:* A refinement of the two previous features, is the entropy that also takes the probabilities of observations at a location L into account. The probability that a user has visited a certain region is defined as the number of observations of the actual users divided by the overall number of observations at this regions. Let $O_{u,L}$ be the observations of a user u at a location L and O_L be all observations at the location L. The probability can then be computed as $prob_L(u) = \frac{|O_{u,L}|}{|O_L|}$. Based on this we can compute the entropy of a certain location L as $E_L = -\sum_{u \in U_L} P_L(u) \cdot log(P_L(u))$ with U_L representing all users observed a the location L. With this definition we computed the mean entropy $R_{E,\mu}(u,v)$ of the locations two users visited in common and the according standard deviation of the mean $R_{E,\sigma}(u,v)$.

V. EXPERIMENTAL SETUP

We conducted different kinds of experiments to study the social interactions between users based on the three different sources of location information.

In order to conduct these experiments we created a network from the social interactions obtained from the online social network of Second Life. In this network, nodes represented the users and edges indicate the social interactions between them. These edges were considered as unweighted and so we add an edge between two users no matter how often they communicated with each other. Further we did not distinguish the actual type of interaction and considered text messages, comments and loves equally. This finally yielded in a network of 152,509 users connected by 270,567 edges. Formally this can be written as $G'_O\langle V'_O, E'_O\rangle$ with V'_O representing the users with an interaction on their feed and $e = (u,v) \in E'_O$ if user u interacted with user v (comment, posting, love). Then we enriched the nodes with the observations $O(u)$ from all three location data sources and removed nodes from the network if this data was not available in all three sources.

Formally this new network can be defined as $G_O\langle V_O, E_O\rangle$ where $V_O = \{u \mid u \in V'_O,\ u \in O_M,\ u \in O_S,\ u \in O_F\}$ and $e = (u, v) \in E_O$ if user u interacted with user v (comment, posting, love). This reduced the network size to 14,508 nodes and 23,446 edges. For the actual experiments we followed Guha *et al.* [12] who suggest to create a balanced set of user-pairs with an interaction and without interaction for the prediction task. In particular we randomly selected 15,000 user-pairs with interaction $\{e^+ = (u, v)|(u, v) \in E_O, u \text{ and } v \in V_O\}$ connecting users V_O^+. The remaining 15,000 edges without interaction in between were created by selecting random user-pairs from the network without interaction $\{e^- = (u, v)|(u, v) \notin E_O, u \text{ and } v \in V_O\}$. Using this network we computed the features described in Section IV for all 30.000 user-pairs and each location source separately. This network-setup implies a baseline of 0.5 for the prediction task in case of random guessing whether a user-pair has interactions or not.

A. Analysis of Homophily

In the first experiment we analyzed similarities and dissimilarities of user-pairs with interaction e^+ and user-pairs without interaction e^- for each location source separately. We computed the mean values of the features and the according standard error in either sources separately. Using a one-sampled Kolmogorov-Smirnov and a Anderson-Darling test showed that none of the distributions of the features was from the family of normal distribution. As a consequence and similarly to Bischoff [3], we compared the variances of all features between interacting and non-interacting user-pairs with a Levene test ($p < 0.01$). To test for significant differences of the means, we employed a Mann-Whitney-Wilcoxon test in case of equal variances and a two-sided Kolmogorov-Smirnov test in case of unequal variances.

B. Feature Engineering

In order to utilize the supervised machine learning algorithms to predict whether or not a user-pair interacted with each other, we had to determine the features that are most suited for this task. To assess the impact of each feature separately we used a simple Collaborative Filtering algorithm for a first rough overview and implemented a method proposed by Liben *et al.* [13]: For every user in the network we created ranked lists of the remaining users in the network based on the homophily obtained from the single features. To evaluate the performance of this approach we compared lists with different length to the actual interaction partners of each user. This was computed as the fraction of correctly identified interaction partners divided by the length of the overall retrieved users also referred to as the *positive predictive value* or *precision*. To validate the results of this approach we additionally employed the built-in Information Gain and the Correlation-Based Feature Subset Selection of the WEKA learning suite [14] to find the most valuable features for supervised learning.

C. Predicting Social Interactions with Supervised Learning

Based on the most valuable features determined for every region source separately, we tried to predict whether two users have a social interaction in the online social network.

TABLE II. FEATURE ENGINEERING WITH COLLABORATIVE FILTERING AND THE ACCORDING INFORMATION GAIN. HIGHLIGHTED FEATURES WERE DERIVED FROM CORRELATION-BASED FEATURE SUBSET SELECTION.

	Features	Info Gain	Collaborative Filtering		
			Pre@5	Pre@10	Pre@20
Monitored Locations	$\mathbf{R_C(u, v)}$	0.048	0.081	0.062	0.048
	$R_{U,\mu}(u, v)$	< 0.01	0.047	0.041	0.039
	$R_{U,\sigma}(u, v)$	< 0.01	0.046	0.040	0.037
	$R_{E,\mu}(u, v)$	< 0.01	0.025	0.029	0.029
	$R_{E,\sigma}(u, v)$	< 0.01	0.046	0.037	0.033
	$R_{F,\mu}(u, v)$	< 0.01	0.047	0.043	0.037
	$R_{F,\sigma}(u, v)$	< 0.01	0.046	0.040	0.035
	$\mathbf{R_{JC}(u, v)}$	0.051	0.071	0.063	0.052
	$\mathbf{R_O(u, v)}$	0.051	0.071	0.063	0.052
	$\mathbf{R_T(u, v)}$	0.012	0.077	0.043	0.023
Shared Locations	$R_C(u, v)$	0.211	0.280	0.252	0.208
	$R_{U,\mu}(u, v)$	< 0.01	0.133	0.119	0.104
	$R_{U,\sigma}(u, v)$	< 0.01	0.185	0.161	0.137
	$R_{E,\mu}(u, v)$	< 0.01	0.122	0.089	0.074
	$R_{E,\sigma}(u, v)$	< 0.01	0.192	0.164	0.129
	$R_{F,\mu}(u, v)$	< 0.01	0.115	0.099	0.091
	$R_{F,\sigma}(u, v)$	< 0.01	0.109	0.108	0.101
	$\mathbf{R_{JC}(u, v)}$	0.208	0.221	0.187	0.157
	$\mathbf{R_O(u, v)}$	0.208	0.221	0.187	0.157
	$\mathbf{R_T(u, v)}$	0.234	0.159	0.121	0.107
Favored Locations	$\mathbf{R_C(u, v)}$	0.040	0.104	0.085	0.060
	$R_{U,\mu}(u, v)$	< 0.01	0.079	0.075	0.055
	$R_{U,\sigma}(u, v)$	< 0.01	0.082	0.074	0.058
	$R_{E,\mu}(u, v)$	< 0.01	0.082	0.076	0.056
	$R_{E,\sigma}(u, v)$	< 0.01	0.086	0.077	0.057
	$R_{F,\mu}(u, v)$	< 0.01	0.081	0.075	0.055
	$R_{F,\sigma}(u, v)$	< 0.01	0.074	0.071	0.056
	$\mathbf{R_{JC}(u, v)}$	0.040	0.108	0.086	0.059
	$R_O(u, v)$	0.040	0.108	0.086	0.059
	$\mathbf{R_T(u, v)}$	0.020	0.002	0.002	0.003

We combined features selected by the Correlation-Based Feature Subset Selection for each location source separately and obtained the three feature sets used for supervised learning algorithms. Due to the split into 15,000 user-pairs with interactions and 15,000 user-pairs without interactions we reduced the experiment to a binary classification problem. To compare the different location-based knowledge sources against each other, we applied the WEKA machine learning suite onto the combined set of features obtained with feature engineering for each region source separately. To do so, we applied three learning algorithms: "Logistic Regression" as it can be easily interpreted, and "Random Forest" and "Support Vector Machine" as both of them are suited for high-dimensional data. For the verification of the results provided by the machine learning tool, we used a ten-fold approach for the split of training set and test set.

VI. RESULTS

In this Section we present the results of the conducted experiments.

A. Analysis of Homophily

We computed the mean values and standard errors for all features of 15,000 user-pairs with interactions and 15,000 user-pairs without interactions in the online social network. Table I summarizes the differences for features applied to all three sources of location-based information.

1) Monitored Locations: On average user-pairs with interaction could be found in 0.5 common regions $R_C(u, v)$, had over 12 total regions $R_T(u, v)$, and Jaccard's Coefficient

$R_{JC}(u,v)$ and observations $R_O(u,v)$ of around 0.05. For user-pairs with interaction we furthermore found an average user count $R_{U,\mu}(u,v)$ of over 179, an entropy $R_{E,\mu}(u,v)$ of 1.52 and a user frequency $R_{F,\mu}(u,v)$ of 637 for commonly visited regions. For user-pairs without interaction we observed less commonly visited regions and total regions as well as Jaccard's Coefficient and observations. In contrast, for features based on the location diversity, i.e. entropy, frequency, and user-count, we observed higher values. With the tests described in Section V we found significant differences for all applied features.

2) Shared Locations: The characteristics of the features applied to the Shared Locations were similar to the features applied to the Monitored Locations. For user-pairs with interaction we observed around 1 common region $R_C(u,v)$, 63 total regions $R_T(u,v)$, and a Jaccard's Coefficient $R_{JC}(u,v)$ and observations $R_O(u,v)$ in the same regions of around 0.03. For common regions we observed a average user-count $R_{U,\mu}(u,v)$ of 22, region entropy $R_{E,\mu}(u,v)$ of 0.8, and region frequency $R_{F,\mu}(u,v)$ of 92. Similar to the Monitored Locations dataset we observed higher values for common regions, Jaccard's Coefficient, observations, and total regions for user-pairs with interaction, whereas frequency, user-count and entropy were lower.

3) Favored Locations: Again we observed similar results as already described for the previous locations dataset but due to the reduced number of picks per user the absolute values were lower. We observed 0.11 common regions $R_C(u,v)$ for users interacting with each other, respectively 0.02 for observations $R_O(u,v)$ and Jaccard's Coefficient $R_{JC}(u,v)$. In contrast these values were nearly 0 for user-pairs without interaction. Interacting users had around 8 total regions $R_T(u,v)$ whereas user-pairs without interaction had only around 7 total regions. For features that model the location diversity ($R_E(u,v)$, $R_F(u,v)$, $R_U(u,v)$) we again observed lower values for users interacting with each other if compared to users without interaction.

B. Feature Engineering

For a rough estimation of the predictability of interactions we employed a Collaborative Filtering algorithm using features applied to the three location-based knowledge sources. Previous results of the analysis of homophily showed that user-pairs with interactions had higher values for common regions, total regions, Jaccard's Coefficient and observations. Hence, we rank this features in this experiment in descending order. Contrary, features based on the location diversity ($R_E(u,v)$, $R_F(u,v)$, $R_U(u,v)$) showed significantly lower values for interacting user-pairs and so we ranked them in ascending order. In addition to Collaborative Filtering, we used WEKA's Information Gain algorithm for verification of these results and finally a Correlation-Based Feature Subset Selection to find valuable features for further prediction. In Table II we present the results of Collaborative Filtering and the according values of the Information Gain algorithm for the features applied to the three location sources.

1) Monitored Locations: The Collaborative Filtering approach unveiled a high predictive power for common regions $R_C(u,v)$, total regions $R_T(u,v)$, respectively Jaccard' Coefficient $R_{JC}(u,v)$ and common observations $R_O(u,v)$ for different list lengths. However, features modeling location diversity like user-count, entropy, frequency of user's common regions performed inferior. This results were inline with the Information Gain algorithm that showed similar results for the computed features. Additionally, Correlation-Based Feature Subset Selection identified these features as most valuable.

TABLE III. PREDICTING INTERACTIONS BETWEEN USER-PAIRS WITH SUPERVISED LEARNING BASED ON COMBINED FEATURES OF DIFFERENT LOCATION SOURCES.

Feature Set	Logistic	SVM	Random Forest
Monitored Locations	**0.632**	0.605	0.618
Shared Locations	**0.849**	0.791	0.846
Favored Locations	**0.630**	0.593	0.628

2) Shared Locations: Collaborative Filtering exposed common region $R_C(u,v)$, Jaccard's Coefficient $R_{JC}(u,v)$, and observations $R_O(u,v)$ as most valuable. These three features plus the total number of regions $R_T(u,v)$ were also identified as best features using the Information Gain algorithm. Similarily, the Correlation-Based Feature Subset Selection algorithm unveiled Jaccard's Coefficient $R_{JC}(u,v)$, common observations $R_O(u,v)$, and the total number of regions $R_T(u,v)$ as the most valuable features in the set.

3) Favored Locations: Similar to the previous result the Collaborative Filtering approach identified the common regions $R_C(u,v)$, Jaccard's Coefficient $R_{JC}(u,v)$ and common observations $R_O(u,v)$ as most valuable. Information Gain additionally puts the total number of regions $R_T(u,v)$ on the list which is also inline with the previous result. Finally, Correlation-Based Feature Subset Selection found common regions $R_C(u,v)$, Jaccard's Coefficient $R_{JC}(u,v)$ and the total number of regions $R_T(u,v)$ to be best suited for further prediction tasks.

C. Predicting Social Interactions

Based on the results of the previous experiment we used the features identified by Correlation-Based Feature Subset Selection for predicting whether two users have an interaction with each other or not. One can find these features highlighted in bold letters in Table II for different region sources. We combined these individual features to feature-sets for every location source separately and predicted the interaction between user-pairs with three different learning algorithms. We utilized *Logistic Regression*, *Support Vector Machine* (SVM), and *Random Forest* and used the Area under the ROC curve (AUC) as main evaluation metric. In Table III the results of these evaluations are shown and one can see that *Logistic Regression* outperforms the two remaining algorithms on each of the three location-datasets. In particular, we found that the feature-set applied to the Shared Location dataset predicted interactions between users with 0.849 AUC which is a boost of +34.9% if compared to baseline for random guessing. For the remaining two region sources we observed a predictability of around 0.63 which is +13% over baseline. Random Forest and SVM showed similar results but performed inferior.

VII. DISCUSSION AND CONCLUSION

In this paper we have harvested data from different sources of the virtual world of Second Life: First we collected social interaction data between users from the online social network *My*

Second Life and second, we collected data from three different and independent location sources, i.e. locations monitored while users were attending events, locations they explicitely share, and their favorite locations. For every single location source we computed 10 features representing the homophily of user-pairs and employed them to predict whether two users had social interaction with each other or not. This section concludes the paper and tries to give answers to the research questions from Section I and provides possible explanations for the results derived from the conducted experiments.

RQ1. To answer the first research question, we evaluated the differences between user-pairs that had an interaction in the online social network and user-pairs without this interaction. This analysis revealed statistically significant differences for nearly all features: User-pairs with interactions on average visited more common regions and had more common observations together. In contrast to this, they visited regions with a lower user-count, frequency, and entropy which can be interpreted as sign of intimacy: Users with interactions already know each other and therefore they meet in places that are less frequented by other users. We could observe this for all three data sources but due to the diverse datasets the characteristics were different: the Shared Locations dataset showed more distinct tendencies than, for instance the picks dataset with the given limit of 10 picks per user.

RQ2. To answer the second research question we employed Collaborative Filtering to predict the social interactions between the users based on 10 different features independently across all location sources. We found that the most valuable features over all the location-based knowledge sources were the number of common regions $R_C(u, v)$, the Jaccard's Coefficient $R_{JC}(u, v)$, and the total number of regions of two users $R_T(u, v)$. Although these characteristics were similar over all sources, we observed differences in the Information Gain. Features applied to the Shared Locations seemed best suited for predicting interactions as the Information Gain was higher if compared to Favored or Monitored Locations.

RQ3. Considering the Information Gain of features applied to the three location sources, we already had the premonition that data obtained from a user's Shared Locations has the highest potential to predict interactions. Indeed, a detailed look at the combined feature sets to predict interactions unveiled that this dataset worked best among all sources. We believe that this is for the following two reasons: First, users can share message from everywhere within the virtual world over their social network and the data collection approach does not miss any data. Second, users explicitly share locations and places they like and spend time in. Other users that visit their profiles because they already know each other, see these locations, and also visit them. This can be seen as an explicit promotion of Shared Locations of a user. We believe that Monitored Location data performed inferior as we only have a clipping of the actual user's visited regions due to limited resources. A similar explanation can be made for the picks data source but here the limiting factor was not the lack of crawling resources but the restriction of 10 picks per user. Overall, the three different learning algorithms applied to the datasets were stable and show similar results over all three sources – Logistic Regression showed the best results whereas Support Vector Machine and Random Forrest were inferior.

For future work we plan to account for the variation of time which we did not consider in this paper.

VIII. ACKNOWLEDGMENTS:

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] 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*. ACM, 2011, pp. 1046–1054.

[2] 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*. ACM, 2010, pp. 119–128.

[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*. ACM, 2012, pp. 47–56.

[4] M. Steurer, C. Trattner, and F. Kappe, "Success factors of events in virtual worlds a case study in second life," in *Network and Systems Support for Games (NetGames), 2012 11th Annual Workshop on*. IEEE, 2012, pp. 1–2.

[5] M. Steurer and C. Trattner, "Predicting interactions in online social networks: an experiment in second life," in *Proceedings of the 4th International Workshop on Modeling Social Media*, ser. MSM '13. New York, NY, USA: ACM, 2013, pp. 5:1–5:8. [Online]. Available: http://doi.acm.org/10.1145/2463656.2463661

[6] ——, "Acquaintance or partner? predicting partnership in online and location-based social networks," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM/IEEE, 2013.

[7] 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*. LAWRENCE ERLBAUM ASSOCIATES LTD, 2003, pp. 519–526.

[8] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," in *SDM06: Workshop on Link Analysis, Counter-terrorism and Security*, 2006.

[9] M. Thelwall, "Homophily in myspace," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 2, pp. 219–231, 2009.

[10] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2011, pp. 1100–1108.

[11] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-spatial properties of online location-based social networks," *Proceedings of ICWSM*, vol. 11, pp. 329–336, 2011.

[12] 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*. ACM, 2004, pp. 403–412.

[13] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2002.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.