Acquaintance or Partner? Predicting Partnership in Online and Location-Based Social Networks

Michael Steurer
Institute for Information Systems and Computer Media
Graz University of Technology
Inffeldgasse 16c, A-8010 Graz
Tel: +43 316 873 5619
Email: michael.steurer@iicm.tugraz.at

Christoph Trattner
Know-Center
Graz University of Technology
Inffeldgasse 13/6, A-8010 Graz
Tel: +43 316 873 30840
Email: ctrattner@know-center.at

***Abstract*—Existing approaches to predicting tie strength between users involve either online social networks or location-based social networks. To date, few studies combined these networks to investigate the intensity of social relations between users. In this paper we analyzed tie strength defined as partners and acquaintances in two domains: a location-based social network and an online social network (Second Life). We compared user pairs in terms of their partnership and found significant differences between partners and acquaintances. Following these observations, we evaluated the social proximity of users via supervised and unsupervised learning algorithms and established that homophilic features were most valuable for the prediction of partnership.**

Keywords*—*Online Social Networks; Location-Based Social Networks; Partner Prediction; Virtual Worlds

I. Introduction

Social networks contain useful information about the relation between their participants and the understanding of their characteristics is a prerequisite to interpret social dynamics [1]. The advent of online social networks afforded large-scale data and topological network features were complemented by homophilic features that model the alikeness of users in a network. Nevertheless, the social proximity between users in the real world is not only driven by online social networks but also by their mobility patterns. The availability of such data changed with the arrival of GPS aware mobile phones and location-based social networks opened a new information source. Not all links in a network are equal and so it is not sufficient to consider them merely as loosely coupled. Granovetter [2] introduced the term "tie strength" and proposed the overlap of neighbors derived from the network topology as an indicator.

A considerable body of research investigating tie strength focused on either online social networks [3], [4], [5] or location-based social networks [6], [7], [8]. However, there are few studies combining both domains are rare. To fill this gap, in this paper we analyze the tie strength between users – defined as *partners* and *acquaintances* - with social proximity features from an online social network and a location-based social network.

Since it is nearly impossible to collect large-scale social network and position data of the same users in a real-world scenario, we obtained datasets for the experiments from the virtual world of Second Life. Text-based interactions between residents (posts, comments, loves) and profile data (affiliated groups, specified interests, partnership information) were harvested from a Facebook-like online social network My Second Life. Additionally, we monitored position and mobility patterns of these users by attending events in the virtual environment over a period of 12 months. To model the social proximity between the users, we computed network topological and homophilic features based on their profiles, as pro- posed in related work.

Overall, our aim was to answer the following research questions:

- *RQ1:* To what extent do partners and acquaintances differ from each other based on social proximity features induced from the online social network and the location-based social network?
- *RQ2:* How well can we predict a partnership between users via the social proximity features from both data sources?
- *RQ3:* Which features offer the highest information gain and the highest accuracy to predict partnership between users?

Based on these questions, we conducted a number of experiments using statistical methods and supervised and unsupervised learning approaches. A statistical analysis of the studied feature showed that significant differences existed between partners and acquaintances. For instance, we discovered that partners were less interested in exploring new locations or meeting new friends than acquaintances and that the number of text interactions and the average spatial distance between users could signify intimate contact. The learning algorithms identified homophilic features, such as attended events and the distance between users, derived from the location-based social network to be most valuable with regard to partnership prediction. Our experiments showed that the combination of features from both domains (=an online social network and a location-based social network) outperformed the features of either domain.

The major contributions of this paper are as follows:

- The introduction of a novel large-scale dataset that incorporates online social network data and location-based social network of the same users.

- The analysis of a large set of social proximity features from an online social network and a location-based social network that allows to predict a partnership between users with an accuracy of 0.933 AUC.

In detail the paper is structured as follows: In Section II we discuss related work. In Section III we shortly introduce the dataset used for our experiments. In Section IV we outline the set of features used for our experiments described in Section V. Section VI presents the results of our study. Finally, Section VII discusses our findings and concludes the paper.

II. Related Work

Relevant related work in this area can broadly be divided into the following two areas: Predicting links and predicting tie strength in online and location-based social networks.

A. Predicting links in online and location-based social networks

Liben-Novell and Kleinberg [9] formalized the problem of predicting new links in a network and developed an approach based on the topology of the network. They used information about direct neighbors and employed the ensemble of all paths from one user to another. This approach yielded in significantly better predictability of new links compared to a random approach. Computationally efficient methods of this structure-centric approach were evaluated by Fire *et al.* [10]. Surprisingly, using only topological features they could successfully find new links that evolve within two hops in the network. However, topological features can only be applied if the structure of the actual network is known. If this is not the case, homophilic features such as a metric for the likeness of two users can be used instead. In their work Towall *et al.* [11] investigated the social network of *MySpace* using homophilic features. Their studies revealed a significant homophily of origin, marital status and the sexual orientation in existing links. Further they uncovered that a friendship in an online social network could even reflect an offline friendship. While these papers were of analytical nature, Mislove *et al.* [12] extended the known attributes of a few users in a network to learn about other users. They used Facebook datasets with educational data and region information as attributes and found that one could infer the attributes of 80% of users from the remaining 20% with an accuracy of 80%. Rowe *et al.* [13] combined topological features and homophilic features in the Chinese microblogging service *Tencent Weibo*. In their work they predict network links and show that homophilic features do not only significantly outperform a random baseline but also topological features.

Scientific work for the link prediction problem was mainly done for online social networks but as more and more position data became available, the combination of both worlds was investigated too. One popular work in this respect is for instance a study of Cranwshaw *et al.* [14] who examined data of the Facebook application *Locaccino* and analyzed the offline mobility data of 489 users. They used the position data, separated it into two categories with topological and homophilic information and tried to predict the online links with the position information. Homophilic data obtained from the location was identified as valuable information but in combination with topological features it performed even better. Scellato *et al.* [15], [16] also revealed the importance of place related features and identified 30% of new links in the *Gowalla* network as place friends and 40% of all links within a range of 100 kilometers. Further they uncovered a weak correlation between the number of friends and their spatial distance. Noulan *et al.* [17] used this fact to predict venues of users in *Gowalla* network. They also achieved best results when combining information from social ties and visited venues.

B. Predicting tie strength in online and location-based social networks

Not all links between nodes in a network are equal and Granovetter [2] tightened the definition by introducing the *tie strength* of connections. Studies by Gilbert *et al.* [4], [3] used Facebook to investigate tie strength between online friends. They combined communication features, topological information and the social distance, and correlated it with 2,000 users who specified their real friends in interviews and questionnaires. They could predict different tie strength with an accuracy of 85% and notably, they were even able to transfer information about the tie strength of two users from one social network to the other, i.e. from Facebook to Twitter.

While the tie strength between users in online social networks was extensively investigated, few studies combined offline networks and tie strength. The first to mention study is by Wang *et al.* [6] who collected the mobile phone data of 6 Million users and measured the tie strength according to the number of calls between user pairs. They found that although new links could be predicted via mobility measures alone, combining them with topological information in the network yielded even better results. With regard to tie strength, they found that its correlation with the user mobility traces and social proximity was weak. This is in agreement with Choi *et al.* [7] who analyzed communication patterns and indoor mobility tracking of 22 office mates. To define tie strength, the users formal and informal contacts were differentiated in their study. Via a supervised learning approach, they could identify links with an accuracy of 85%. Finally, the last to mention paper in this respect is the work of Bischoff [8]. In her work the author studied the social network *Last.fm* in respect to the geo location of user and attended events. Tie strength was defined as the number of commonly attended events, and online communications and music tastes were used to predict it. The results were in agreement with previous works, confirming a correlation between tie strength and data from the online social network.

III. Datasets

We conducted our experiments using online social network data and location-based social network data obtained from the virtual world of Second Life. There were several reasons for choosing Second Life online social network. First, unlike such networks as Facebook, My Second Life does not restrict extensive crawling of the users profiles. Second, in contrast to real-world online social networks, most of the profiles in My Second Life are public, i.e., a large fraction of the network can be mined. Third, although in virtual worlds the user location information can be harvested automatically, in the real world it is nearly impossible to obtain large-scale user tracking data. In this Section we describe the data collection process for our experiments.

A. Location-Based Social Network Dataset

Similar to the real life, residents of Second Life can host events and announce them to the public. Users log into the Second Life Web page and create new events with *name*, *description*, *location* and *start time* and assign them to one out of ten predefined categories, *e.g. Nightlife* or *Live Music*. Further, events have three maturity ratings that depend on the rating of the location: *General* without any age restrictions, *Mature* with users at least 16 years old, and finally *Adult* accessible only for grown-up users.

One of the many advantages of using Second Life as testbed for our experiments is the fact that all events are announced to the general public on the Second Life website – called the Second Life event calendar. In order to harvest this kind of data we implemented a simple Web-based crawler which collected all relevant event information on a daily basis. Starting with March 2012, we collected 262,234 events over a period of 12 months [18].

To participate in the virtual world, users register with Second Life, download the client software from Linden Labs and log in. Among other third party clients, *libopenmetaverse*[1] is an open-source client for the command-line to enter the environment. It can be run as a server process and the functionality can be easily enhanced due to the modular design. We added new capabilities to automatically move around in the virtual world and collect information of surrounding users. These user-bots were controlled by a centralized server-instance that sent them to places with ongoing events. On average a bot needed 1 minute to move to a new location and collect the position data of surrounding users. To speed up the collection process and to visit more events concurrently, we employed a pool of 15 bots that alternately visited events. The collected information comprised user names, accurate position of the observed users, and a time stamp. Overall, we collected nearly 19 million data samples of 410,619 different users in 4,105 different locations.

The naive approach to create a location-based social network out of this huge amount of data would have been to interlink two users with each other whenever they met. Since this would yield in a network with billions of edges, we applied a simple heuristic to prune our data. In particular, we only interlinked two users with each other in the network if they were seen concurrently in the same region on two different days. With this simple approach at hand we were able to reduce the number of edges to 4,473,739. Formally we define this network as graph $G_L\langle V_L, E_L\rangle$ with V_L representing the users in the network and $e = (u, v) \in E_L$ if users u and v were concurrently observed in the same region on two different days. In Table I we present the basic properties of the network.

B. Online Social Network Data

In 2007 Linden Labs introduced the online social network platform *My Second Life* which is similar to Facebook or Google+. The target group are residents of Second Life to share text messages, comments or loves (similar to Facebook's "Likes"). Users of Second Life automatically have a profile without additional registration and by default these profiles are opened for public access. In contrast to Facebook, there is no mutual friendship confirmation between users and every user can post onto the Feed of each other (similarly to Facebook's "Wall") without their explicite permission. Besides the interactions with others, users can enhance their profiles and describe themselves with a biography, interests, and their partnership status. Users can even marry in Second Life but a wedding is not free of charge and costs 10 Linden Dollars (Linden Dollar is the virtual currency used in Second Life - 1 US Dollar equals approximately 258 Linden Dollars). Though, nothing is forever and canceling this partnership costs 25 Linden Dollars.

TABLE I. BASIC METRICS OF THE TWO NETWORKS AND THEIR COMBINATION USED FOR THE EXPERIMENTS.

Name	Location-Based G_L	Online G_O	$G_L + G_O$
Type	undirected	directed	directed
#Nodes	$156,844$	$152,509$	$44,603$
#Edges	$4,473,739$	$270,567$	$1,419,543$
Degree	57.05	3.54	63.65

To harvest this data, we extracted the set of user names from the position dataset and attempted to download their interaction data and profile information with groups, interests, and partner. We extracted the user names of the interaction partners and fetched the missing ones iteratively until no new users were found any more. Overall, we downloaded the profile data of 152,509 users with interactions on their walls and identified 1,084,002 postings, 459,734 comments, 1,631,568 loves and 285,528 unique groups. On average users joined 15.61 groups specified 6.5 interests. 39,936 users were in a partnership which resulted in 18,468 couples in the whole dataset. Formally, this network is defined as $G_O = \langle V_O, E_O\rangle$, with V_O representing the users with interactions on their Feed, and $e = (u, v) \in E_O$ if user u interacted with user v (posting, comment, love). In Table I we present the basic properties of the network.

IV. Feature Description

As already outlined in the introductory part of this paper, it is our aim to study the extent to which partnership between users can be predicted based on two different types of data – online social network and location-based social network data. To that end, we induce two different types of feature sets from our data sources: network topological and homophilic features [13], [19].

A. Location-Based Social Network Features

1) Topological Features: Users in online networks with small-world characteristics are clustered locally and the more neighbors two users have in common, the closer they are connected. With the formal definition of the neighbors of a node $u \in V_L$ as $\Theta(u) = \{v \mid (u, v) \in E_L\}$ this feature could be computed as $L_{CN}(u, v) = |\Theta(u) \cap \Theta(v)|$. This measure indicated the overlap of neighbors regardless of the total number of neighbors the users have. To take this into account, we computed *Jaccard's Coefficient* as the number of

[1] http://lib.openmetaverse.org/

common neighbors divided by the total number of neighbors of two users: $L_{JC}(u, v) = \frac{|\Theta(u) \cap \Theta(v)|}{|\Theta(u) \cup \Theta(v)|}$.

A refinement of this metric was proposed by Adamic-Adar [20]. As not all neighbors in a network have the same tie strength, they added weights to the links and computed the relation between two users as $L_{AA}(u, v) = \sum\limits_{z \in \Theta(u) \cap \Theta(v)} \frac{1}{log(|\Theta(z)|)}$.

Another feature to measure the structural overlap of two users was introduced by Cranshaw *et al.* [14]. They introduced the "neigbourhood overlap" as the number of common neighbors divided by the sum of neighbors of either users. Formally, this can be written as $L_{NO}(u, v) = \frac{|\Theta(u) \cap \Theta(v)|}{|\Theta(u) + \Theta(v)|}$.

Active users within a network are more likely to form new interactions than users with less activity. "Preferential Attachment Score" was first mention by Barabasi *et al.* [21] and is the product of the sum of neighbors of either users: $L_{PA}(u, v) = |\Theta(u)| \cdot |\Theta(v)|$.

2) Homophilic Features: In contrast to the topological features in the previous section, homophilic features directly represent the alikeness of user-pairs. These features do not dependent on their direct neighbors in the network or the structure of the network per se because they are only based on properties of either nodes. These features have been identified as valuable resource for link prediction in several studies [12], [16], or [11].

As outlined before, we implemented user-bots that monitored present users at event sites. Using the position data of users, respectively the location and time span of events, we identified all events a user u visited over a year: $\Pi(u) = \{e_1, \ldots, e_n\}$ where e_i represented the i'th event out of n visited. With this simple metric we computed the number of events two users attended in common $E_C = |\Pi(u) \cap \Pi(v)|$, the total number of events $E_T = |\Pi(u) \cup \Pi(v)|$, and finally their fraction $E_{JC} = \frac{|\Pi(u) \cap \Pi(v)|}{|\Pi(u) \cup \Pi(v)|}$.

A refinement of this measure also takes the trajectory of the visited events into account. For each user pair (u, v) we created two vectors $\boldsymbol{\epsilon}(u)$, $\boldsymbol{\epsilon}(v)$ that represent their totally visited events. The j'th component of each vector $\boldsymbol{\epsilon}$ was set to 1 if the user visited the actual event and was set to 0 if it did not. Then we computed the cosine similarity of these two vectors which is formally defined as $E_{CS} = \frac{\boldsymbol{\epsilon}(u) \cdot \boldsymbol{\epsilon}(v)}{||\boldsymbol{\epsilon}(u)|| \cdot ||\boldsymbol{\epsilon}(v)||}$ where $||\boldsymbol{\epsilon}||$ represented the Euclidean length of the vector.

Similar measures can be based on the categories and the maturity rating of events. Events are assigned to different categories and for each user u we created a vector $\boldsymbol{\delta}$ of length, where every item represented the number of events attended in a category. We computed the cosine similarity of two users' vectors $\boldsymbol{\delta}(u)$ and $\boldsymbol{\delta}(v)$ as $E_{CCos} = \frac{\boldsymbol{\delta}(u) \cdot \boldsymbol{\delta}(v)}{||\boldsymbol{\delta}(u)|| \cdot ||\boldsymbol{\delta}(v)||}$. The same measure was applied to the maturity rating of events: $E_{MCos} = \frac{\boldsymbol{\gamma}(u) \cdot \boldsymbol{\gamma}(v)}{||\boldsymbol{\gamma}(u)|| \cdot ||\boldsymbol{\gamma}(v)||}$ with $\boldsymbol{\gamma}$ representing number of events a users visited with the according maturity level.

TABLE II. AREA UNDER THE ROC CURVE (AUC) FOR PREDICTING PARTNERSHIP WITH DIFFERENT FEATURE SETS AND LEARNING ALGORITHMS. THE BEST ALGORITHM FOR EACH FEATURE SET IS HIGHLIGHTED IN BOLD LETTERS.

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

Besides the different category and maturity setting of events, we formally defined the event locations a user u visited over the observation time span as $P(u)$. According to this, we computed the similarity of two users with respect to their visited regions with already described approaches. First, we measured the number of common regions two users visited, not necessarily at the same time $R_C(u, v) = |P(u) \cap P(v)|$, second the total regions of two users $R_T(u, v) = |P(u) \cup P(v)|$ and finally Jaccard's Coefficient as a combination of both: $R_{JC} = \frac{|P(u) \cap P(v)|}{|P(u) \cup P(v)|}$.

Finally, we present two features that reflected the user's activity. First, we extracted the number of days two users were concurrently seen in the same region $A_{DS}(u, v)$ and second, we defined the average distance between them: with the accurate position of every user, we computed the Euclidean distance between them and averaged over all observations to get the spatial proximity $A_D(u, v)$ of the users u and v.

B. Online Social Network Features

1) Topological Features: The topological features described in this section are similar to the topological features in the location-based social network. The only difference is that the online social network contains more hidden information due to its directed structure. We defined the neighbors of a user u in the network with respect to the communication direction: neighbors that received a message from user u were denoted as $\Gamma(u)^+ = \{v \mid (u, v) \in E_O\}$ and neighbors that send messages to user u as $\Gamma(u)^- = \{v \mid (v, u) \in E_O\}$.

The first and most simple measure was the number of common friends a pair of users had. Due to the different definitions of neighbors, we defined the common outgoing neighbors as $O^+_{CN}(u, v) = |\Gamma^+(u) \cap \Gamma^+(v)|$ and the common incoming neighbors as $O^-_{CN}(u, v) = |\Gamma^-(u) \cap \Gamma^-(v)|$.

The relation between common friends of two users and their total friends is Jaccard's Coefficient and could be seen as a measure for exclusiveness of this relation. Again, we split it into two features $O^+_{JC}(u, v) = \frac{|\Gamma^+(u) \cap \Gamma^+(v)|}{|\Gamma^+(u) \cup \Gamma^+(v)|}$ and $O^-_{JC}(u, v) = \frac{|\Gamma^-(u) \cap \Gamma^-(v)|}{|\Gamma^-(u) \cup \Gamma^-(v)|}$.

In their paper Cheng *et al.* [22] investigate in the reciprocity of user communication in a directed network and to take this bidirectional communication into account, we computed $O_R(u, v) = 1$ if $(u, v) \in E_O, (v, u) \in E_O$ and $O_R(u, v) = 0$ if $(u, v) \in E_O, (v, u) \notin E_O$. Furthermore they proposed a modification to the Adamic-Adar measure for directed networks which can be written as $O^-_{AA}(u, v) = \sum\limits_{z \in \Gamma^-(u) \cap \Gamma^-(v)} \frac{1}{log(|\Gamma^-(z)|)}$.

"Preferential Attachment Score" takes the level of activity into account and due to the directed structure used

TABLE III. THE MEAN VALUES AND STANDARD ERRORS FOR TOPOLOGICAL AND HOMOPHILIC FEATURES IN THE ONLINE SOCIAL NETWORK. THE SIGNIFICANCE FOR THE DIFFERENCES OF PARTNERS AND ACQUAINTANCES IS $^{***}p < 0.001$.

Features			Partners	Acquaintances
Online Social Network	Topological	$O^{+}_{CN}(u,v)^{***}$	1.08 ± 0.23	11.93 ± 0.13
		$O^{-}_{CN}(u,v)^{***}$	1.12 ± 0.18	11.70 ± 0.13
		$O^{+}_{JC}(u,v)^{***}$	0.03 ± 0.00	0.05 ± 0.00
		$O^{-}_{JC}(u,v)^{***}$	0.06 ± 0.00	0.05 ± 0.00
		$O_{AA}(u,v)^{***}$	0.81 ± 0.10	6.30 ± 0.07
		$O^{+}_{PS}(u,v)^{***}$	361.17 ± 107.91	6921.79 ± 115.11
		$O^{-}_{PS}(u,v)^{***}$	367.21 ± 107.91	9854.15 ± 132.40
		$O_{RE}(u,v)^{***}$	0.49 ± 0.01	0.29 ± 0.00
	Homophilic	$G_C(u,v)^{***}$	2.30 ± 0.09	0.50 ± 0.01
		$G_{JC}(u,v)^{***}$	0.06 ± 0.00	0.01 ± 0.00
		$I_C(u,v)$	0.05 ± 0.01	0.06 ± 0.00
		$I_{JC}(u,v)$	0.00 ± 0.00	0.00 ± 0.00
		$P_A(u,v)^{***}$	27.25 ± 0.67	18.53 ± 0.13
		$P_C(u,v)^{***}$	5.00 ± 0.50	2.02 ± 0.07
		$P_I(u,v)^{***}$	19.40 ± 1.81	13.11 ± 0.29
		$P_L(u,v)^{***}$	12.33 ± 1.35	10.47 ± 0.24
		$P_P(u,v)^{***}$	2.07 ± 0.13	0.62 ± 0.06

TABLE IV. THE MEAN VALUES AND STANDARD ERRORS FOR TOPOLOGICAL AND HOMOPHILIC FEATURES IN THE LOCATION-BASED SOCIAL NETWORK. THE SIGNIFICANCE FOR THE DIFFERENCES OF PARTNERS AND ACQUAINTANCES IS $^{***}p < 0.001$.

Features			Partners	Acquaintances
Location-Based Social Network	Topological	$L_{CN}(u,v)^{***}$	52.30 ± 5.48	53.33 ± 1.19
		$L_{JC}(u,v)^{***}$	0.29 ± 0.01	0.17 ± 0.00
		$L_{AA}(u,v)^{***}$	77.87 ± 6.03	181.24 ± 3.15
		$L_{PS}(u,v)^{***}$	92324.84 ± 42759.98	82591.90 ± 3771.87
		$L_{NO}(u,v)^{***}$	0.81 ± 0.00	0.87 ± 0.00
	Homophilic	$E_C(u,v)^{***}$	9.51 ± 0.85	1.00 ± 0.02
		$E_T(u,v)^{***}$	32.22 ± 1.60	41.45 ± 0.27
		$E_{JC}(u,v)^{***}$	0.31 ± 0.01	0.02 ± 0.00
		$E_{Cos}(u,v)^{***}$	0.43 ± 0.01	0.04 ± 0.00
		$E_{CCos}(u,v)^{***}$	0.82 ± 0.01	0.66 ± 0.00
		$E_{MCos}(u,v)^{***}$	0.76 ± 0.01	0.19 ± 0.00
		$R_C(u,v)^{***}$	4.63 ± 0.12	3.12 ± 0.04
		$R_T(u,v)^{***}$	10.84 ± 0.24	16.95 ± 0.20
		$R_O(u,v)^{***}$	0.31 ± 0.00	0.18 ± 0.00
		$A_S(u,v)^{***}$	11.54 ± 0.44	6.74 ± 0.11
		$A_D(u,v)^{***}$	5.02 ± 0.27	11.70 ± 0.22

$O^{+}_{PS}(u,v) = |\Gamma^{+}(u) \cdot \Gamma^{+}(v)|$, and one for received-message neighbors $O^{-}_{PS}(u,v) = |\Gamma^{-}(u) \cdot \Gamma^{-}(v)|$.

2) Homophilic Features: Users of Second Life can join groups and specify interests on their profiles to state their opinions. The structure of the data is quite similar for interests and groups, so we could apply the same mechanisms to indicate the similarity between a pair of users. Formally, we defined the groups of a user u as $\Delta(u)$ and the specified interests as $\Psi(u)$. For each pair of users in the network we defined the common interests and the common groups they share: $G_C(u,v) = |\Delta(u) \cap \Delta(v)|$, respectively $I_C(u,v) = |\Psi(u) \cap \Psi(v)|$. Further, we computed Jaccard's Coefficient to take the total number of groups and interests into account: $G_{JC}(u,v) = \frac{|\Delta(u) \cap \Delta(v)|}{|\Delta(u) \cup \Delta(v)|}$, respectively $I_{JC}(u,v) = \frac{|\Psi(u) \cap \Psi(v)|}{|\Psi(u) \cup \Psi(v)|}$.

Further, users can share text messages, comments, or loves with others and the intensity of this communication could be an indicator of their partnership. As a consequence we measured the number of occurrences for each type of interaction and summed it up for the overall number of interactions between users. We defined $P_P(u,v)$ as the number of text messages, $P_C(u,v)$ as the number of comments, $P_L(u,v)$ as the number of loves, and $P_I(u,v) = P_P(u,v) + P_C(u,v) + P_L(u,v)$ as the number of interactions between user u and v.

Another measure for the proximity of users is the average message length of all interactions between them. Hence, we computed the average message length $P_A(u,v)$ as concatenation of all postings and comments between user u and user v and divided it by their quantity.

V. EXPERIMENTAL SETUP

In the previous section we created two networks derived from two different domains, and described topological and homophilic features for both. In this section we present the experiments to answer the research questions. First, we describe the analysis to compare partners and acquaintances upon their features to determine significant differences. Then we show supervised and unsupervised learning approaches to evaluate these features regarding their predictability of partnership.

To conduct all further experiments, we merged the online and the location-based social network into one mixed network $G\langle V, E\rangle$. This new network comprised of users u that can be found in the directed online social network G_O as well as in the undirected location-based social network G_L: $V = \{u \mid u \in V_O,\ u \in V_L\}$. The edges E representing the relations between these users were defined as the union of edges from either networks: $E = \{(u,v) \mid (u,v) \in E_O$ or $(u,v) \in E_L$, and $u, v \in V\}$.

The overall number of users in this network was 44,603 and the number of edges was 1,419,543 with 1,584 user pairs in a partnership (see Table I for basic characteristics of the network).

A. Comparing Partners and Acquaintances

To answer the first research question, we analyzed the similarities and dissimilarities between partners and acquaintances with respect to the features described in Section IV. We split the user pairs into balanced sets of partners and acquaintances, and computed mean values and standard errors of all features in either sets separately. The one-sampled Kolmogorov-Smirnov and the Anderson-Darling test showed that none of the distributions of the features were from the family of normal distribution. As a consequence and similarly to Bischoff [8], we compared the variances of all features between partners and acquaintances using a Levene test ($p < 0.01$). To test significant differences of mean values, we employed Mann-Whitney-Wilcoxon test in case of equal variances and a two-sided Kolmogorov-Smirnov test in case of unequal variances.

B. Predicting Partnership

Residents of Second Life can marry their friends and the partnership information with the partner's name appears on their profiles. To answer the remaining research questions, we employed the social proximity features to predict whether a user pair is in a partnership or not.

Basically, we used two different techniques:

- *Predicting Partnership with Supervised Learning:* In this approach we applied different learning algorithms onto a training set to identify characteristics of partnership and then verified this in a test set. To do so, we reduced the prediction problem to a binary classification problem by selecting 1,500 partners and

TABLE V. RESULTS OF THE SUPERVISED AND UNSUPERVISED APPROACH TO PREDICT THE PARTNERSHIP BETWEEN USERS. HIGHLIGHTED FEATURES HAVE AN INFORMATION GAIN > 0.1.

	Features	Supervised		Unsupervised		
		AUC	Gain	SR@1	SR@5	SR@10
Online Social Network	$\mathbf{O^{+}_{CN}(u,v)}$	**0.648**	**0.104**	**0.153**	**0.329**	**0.505**
	$O^{-}_{CN}(u,v)$	0.598	< 0.1	0.120	0.329	0.593
	$\mathbf{O^{+}_{JC}(u,v)}$	**0.629**	**0.105**	**0.186**	**0.428**	**0.604**
	$O^{-}_{JC}(u,v)$	0.445	< 0.1	0.186	0.461	0.637
	$O_{AA}(u,v)$	0.604	< 0.1	0.120	0.329	0.549
	$\mathbf{O^{+}_{PS}(u,v)}$	**0.654**	**0.160**	**0.033**	**0.230**	**0.439**
	$\mathbf{O^{-}_{PS}(u,v)}$	**0.744**	**0.296**	**0.044**	**0.186**	**0.450**
	$O^{+}_{R}(u,v)$	0.579	< 0.1	0.120	0.384	0.637
	$G_C(u,v)$	0.594	< 0.1	0.252	0.472	0.604
	$\mathbf{G_{JC}(u,v)}$	**0.603**	**0.134**	**0.296**	**0.472**	**0.604**
	$I_C(u,v)$	0.508	< 0.1	0.076	0.296	0.538
	$I_{JC}(u,v)$	0.508	< 0.1	0.076	0.296	0.538
	$P_A(u,v)$	0.642	< 0.1	0.076	0.505	0.725
	$P_C(u,v)$	0.542	< 0.1	0.065	0.373	0.538
	$P_I(u,v)$	0.576	< 0.1	0.054	0.241	0.472
	$P_L(u,v)$	0.617	< 0.1	0.022	0.164	0.384
	$P_P(u,v)$	0.615	< 0.1	0.120	0.362	0.549
Location-based Social Network	$L_{CN}(u,v)$	0.510	< 0.1	0.384	0.626	0.703
	$L_{JC}(u,v)$	0.499	< 0.1	0.417	0.615	0.703
	$L_{AA}(u,v)$	0.760	< 0.1	0.230	0.626	0.681
	$L_{PS}(u,v)$	0.693	< 0.1	0.241	0.626	0.681
	$L_{NO}(u,v)$	0.520	< 0.1	0.351	0.571	0.659
	$\mathbf{E_C(u,v)}$	**0.821**	**0.294**	**0.483**	**0.736**	**0.791**
	$E_T(u,v)$	0.615	< 0.1	0.000	0.131	0.406
	$\mathbf{E_{JC}(u,v)}$	**0.852**	**0.372**	**0.549**	**0.736**	**0.802**
	$\mathbf{E_{Cos}(u,v)}$	**0.854**	**0.378**	**0.538**	**0.747**	**0.802**
	$E_{CCos}(u,v)$	0.672	< 0.1	0.175	0.417	0.604
	$\mathbf{E_{MCos}(u,v)}$	**0.788**	**0.237**	**0.307**	**0.560**	**0.703**
	$R_C(u,v)$	0.376	< 0.1	0.516	0.681	0.692
	$R_T(u,v)$	0.685	< 0.1	0.230	0.582	0.659
	$R_O(u,v)$	0.579	< 0.1	0.549	0.670	0.692
	$A_S(u,v)$	0.343	< 0.1	0.461	0.681	0.703
	$A_D(u,v)$	0.743	< 0.1	0.197	0.582	0.659

acquaintances from the network whose proximity features were fed into the WEKA machine learning suite [23]. To validate the obtained results we used a ten-fold cross validation approach.

- *Predicting Partnership with Unsupervised Learning:* Due to the balanced data set of partners and acquaintances, the binary classification problem has a baseline of 0.5 when randomly guessing. However, to better estimation the performance and importance of the supervised learning algorithm features, it is recommended to compare the results with an unsupervised learning approach [8]. For that purpose, we used a simple Collaborative Filtering technique that was first proposed by Liben *et al.* [9]: For every user in a partnership, we rank all acquaintances according to the features described in Section IV. Next, we ranked potential partners for every feature separately and computed the success rate of finding the partner within a results list of length k.

VI. RESULTS

This section presents the results of the conducted experiments.

A. Comparing Partners and Acquaintances

We computed the mean values and standard errors for all features of partners and acquaintances and used the Mann-Whitney-Wilcoxon test, respectively Kolmogorov-Smirnov test to determine whether they differ significantly. In Table III and IV we present the differences between partners and acquaintances for features from the online social network and the location-based social network.

1) Online Social Network Features: At first glance, Table III reveals that partners were less connected in the network than acquaintances. In particular, we can see that acquaintances had approximately 11 common interaction partners $O^{+}_{CN}(u,v)$, $O^{-}_{CN}(u,v)$ whereas partners had about 1 partner in common. Similar observations were made for Jaccard's Coefficient $O^{+}_{JC}(u,v)$, $O^{-}_{JC}(u,v)$, Adamic-Adar $O_{AA}(u,v)$, and Preferential Attachment Score $O^{+}_{PS}(u,v)$, $O^{-}_{PS}(u,v)$. For the communication direction $O_{RE}(u,v)$ we examined bidirectional communication in nearly 50% of all partnerships but in only 30% of all acquaintances. All topological features were significantly different.

Although the topological features would let us assume that partners did not actively participate in the online social network, the interaction data drew a different picture: on average partners had 19.40 interactions $P_I(u,v)$ which was significantly more than acquaintances with 13.11. This significant difference was observed for postings, comments, and loves as well. Accordingly, we noticed an average message length $P_A(u,v)$ of 27.25 characters per message for partners but only 18.53 characters for acquaintances. On average partners had 2.30 common groups $G_C(u,v)$ versus 0.50 for acquaintances but in contrast none of the interest-based features $I(u,v)$ was meaningful due to small values and insignificant differences.

2) Location-Based Social Network Features: Topological features of the location-based social network revealed similar characteristics as topological features of the online social network. With a significant difference we observed over 52 common neighbors $L_{CN}(u,v)$ for partners and over 53 common neighbors for acquaintances. The results of the Adamic-Adar measure $L_{AA}(u,v)$ and Preferential Attachment Score $L_{PS}(u,v)$ go in line but Jaccard's Coefficient $L_{JC}(u,v)$ was slightly higher for partners.

The overlap of visited regions $R_C(u,v)$ with 4.63 was significantly higher than the according feature of acquaintances with 3.12 but interestingly, the opposite was observed for the total number of regions $R_T(u,v)$. Similar results were discovered for the common and total number of events $E_C(u,v)$, $E_T(u,v)$. Partners met each other on over 11 days compared to over 6 days of acquaintances and during their co-occurrence they had a significantly less spatial distance (5.02 vs. 11.70 meter) between them.

B. Predicting Partnership with Supervised Learning

To evaluate the performance of our feature sets from either sources we utilized popular supervised learning approaches such as *J.48*, *Logistic Regression* and *Support Vector Machine* (*SVM*). We used the area under the ROC curve (AUC) as our main evaluation metric [24], [25]. As shown in Table II, we find that *Logistic Regression* outperformed the remaining algorithms in all feature sets except the topological features of the online social network. This one-off feature was neglected because it did not influence the overall result when all features were used. The combination of features from the online social network yielded in a predictability of partnership of 0.878 AUC but the combination of features from the location-based social network even outperformed this result with 0.905 AUC. The combination of features from both networks resulted in 0.933 AUC and therefore outperformed the baseline by +43.3%.

To further determine the usefulness of each of our features separately we evaluated the predictive power of each of our features separately (see below) with the Logistic Regression algorithm and determined their information gain using WEKA's attribute selection algorithm. The results of these computations are presented in Table V.

1) Online Social Network Features: Preferential Attachment Scores for neighbors who sent message $O^{-}_{PS}(u, v)$ in the online social network had the highest information gain with 0.296 and corresponding prediction factor of 0.744 AUC. For homophilic features, Jaccard's Coefficient for groups $G_{JC}(u, v)$ was around 0.6 AUC and features based on the interests of users $I_C(u, v)$, $I_{JC}(u, v)$ did not work at all. Communication based features with number of postings and loves, respectively average message length could predict a partnership with about 0.6 AUC.

2) Location-Based Social Network Features: Interestingly, none of the topological based features in the location-based social network had an information gain over 0.1 and they were outperformed by event-based homophilic features. Jaccard's Coefficient $E_{JC}(u, v)$ and the cosine similarity $E_{JC}(u, v)$ of events had the highest information gain and correctly predicted partnership with about 0.85 AUC. The average distance between two avatars $A_D(u, v)$ had a predictive power of 0.743 AUC. Further, we observed a remarkable low AUC for the days two avatars were concurrently seen in the same regions $A_D(u, v)$ (0.343) and the number of common regions $R_C(u, v)$ (0.376).

C. Predicting Partnership with Unsupervised Learning

Additionally, we compared the results of the supervised prediction algorithm with the outcome of an unsupervised learning algorithm. This is useful to better estimate performance in real applications [8] and to support our previous findings. Hence, we implemented a simple Collaborative Filtering approach to rank potential partners of users according to their similarity. The success rates that the actual partner was found in lists of length 1 (SR@1), 5 (SR@5), and 10 (SR@10) are presented in Table V.

Obviously, we could observe an increasing hit rate with increasing number of suggested users, i.e. increasing list length. In group related features from the online social network, 29.6% of all partners were ranked on top of the list. In the domain of the location-based social data, event related features performed best and Jaccard's Distance and the Cosine Similarity were identified as most valuable features. In over 53% of all cases the partner of a user was ranked on the top-position in the list.

VII. 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 evaluated the differences between partners and acquaintances in the online social network and the location-based social network. Interestingly, this analysis revealed that partners had less common neighbors and communication partners than acquaintances in the location-based social network and the online social network. Contrary to this observation, homophilic features revealed a strong affection between partners; we found evidence that partners shared more common groups, had more interactions between them and attended more events together. Besides the event features, we also identified a higher number of jointly visited regions for partners but overall a lesser number of total regions. We interpret this small number of total regions combined with low overlap of neighbors in the network as a sign for intimacy. Users in a partnership are familiar with their environment and are not anxious to meet new users in unknown places. This is in line with the observation that partners were on average spatially closer than acquaintances during co-occurrence.

For the second research question *RQ2*, we predicted the partnership between users and merged the networks and the according features into one network. We reduced the prediction problem to a binary classification problem and evaluated our features using three different learning algorithms. Although all of them showed similar characteristics, *Logistic Regression* performed best which goes along with related work in this area [26], [13]. Homophilic features were approved as valuable source for the prediction of partnership in both domains. This result can be compared to the real world where the alikeness of two users, i.e. homophily, is a premise for a working partnership. Furthermore, the combination of features from two different sources yielded in a significant improvement of predictability compared to either sources alone. For our experiment, we finally achieved a predictability of partnership of 0.933 AUC.

Finally, to answer the third research question *RQ3*, we compared the predictive power of single features with a simple Collaborative Filtering approach. To that end, we computed the predictability of partnership for every feature with Logistic Regression and ranked lists of users' similarity. The predictive power of topological features performs well for supervised learning and badly for unsupervised learning. In contrast, homophilic features of either datasets have a high predictive power with both concepts. This lets us assume that homophilic features have a better correlation for tie strength than topological features in general. In particular we identified features derived from the attitude of users, like events and groups, as features with the highest information gain. Further, interpersonal bonding with spatial distance and number of postings were detected as evidence for a partnership between two users.

Our results can be summarized as follows:

- We collected data of the over 44,000 users with activity in two different networks: an *online social network* and a *location-base social network*.
- Analyzing topological and homophilic features in these networks revealed significant differences between *partners* and *acquaintances*.
- We identified *homophilic features* of the location-based network as most valuable to predict a partnership between users.

In the future, we plan to investigate the mobility patterns of users and consider their activities in terms of time. We hope to further improve our results and obtain a deeper insight into the relationships of users in online social networks and location-based social networks.

ACKNOWLEDGMENT

This work is supported by the Know-Center and the EU funded project Learning Layers. The Learning Layers project is supported by the European Commission within the 7th Framework Programme under Grant Agreement 318209, under the DG Information society and Media (E3), unit of Cultural heritage and technology-enhanced learning. 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] J. S. Coleman, "Social capital in the creation of human capital," *American journal of sociology*, pp. S95–S120, 1988.

[2] M. S. Granovetter, "The strength of weak ties," *American journal of sociology*, pp. 1360–1380, 1973.

[3] E. Gilbert, "Predicting tie strength in a new medium," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 2012, pp. 1047–1056.

[4] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009, pp. 211–220.

[5] E. Zheleva, L. Getoor, J. Golbeck, and U. Kuter, "Using friendship ties and family circles for link prediction," in *Advances in Social Network Mining and Analysis*. Springer, 2010, pp. 97–113.

[6] 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.

[7] J. Choi, S. Heo, J. Han, G. Lee, and J. Song, "Mining social relationship types in an organization using communication patterns," in *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 2013, pp. 295–302.

[8] 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.

[9] 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.

[10] 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)*. IEEE, 2011, pp. 73–80.

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

[12] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, "You are who you know: inferring user profiles in online social networks," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 251–260.

[13] M. Rowe, M. Stankovic, and H. Alani, "Who will follow whom? exploiting semantics for link prediction in attention-information networks," in *Proceedings of the 11th international conference on The Semantic Web - Volume Part I*, ser. ISWC'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 476–491.

[14] 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.

[15] 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.

[16] 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.

[17] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, "Mining User Mobility Features for Next Place Prediction in Location-based Services," in *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 2012, pp. 1038–1043.

[18] M. Steurer, C. Trattner, and F. Kappe, "Success factors of events in virtual worlds a case study in second life," in *NetGames*, 2012, pp. 1–2.

[19] 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.

[20] L. Adamic and E. Adar, "Friends and neighbors on the web," *Social networks*, vol. 25, no. 3, pp. 211–230, 2003.

[21] A. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[22] 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)*. IEEE, 2011, pp. 49–56.

[23] 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.

[24] 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.

[25] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, no. 3, pp. 299–310, 2005.

[26] 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*. ACM, 2010, pp. 641–650.