Exploiting Place Features in Link Prediction on Location-based Social Networks

Salvatore Scellato
Computer Laboratory
University of Cambridge
ss824@cam.ac.uk

Anastasios Noulas
Computer Laboratory
University of Cambridge
an346@cam.ac.uk

Cecilia Mascolo
Computer Laboratory
University of Cambridge
cm542@cam.ac.uk

ABSTRACT

Link prediction systems have been largely adopted to recommend new friends in online social networks using data about social interactions. With the soaring adoption of location-based social services it becomes possible to take advantage of an additional source of information: the places people visit.

In this paper we study the problem of designing a link prediction system for online location-based social networks. We have gathered extensive data about one of these services, Gowalla, with periodic snapshots to capture its temporal evolution. We study the link prediction space, finding that about 30% of new links are added among "place-friends", i.e., among users who visit the same places. We show how this prediction space can be made 15 times smaller, while still 66% of future connections can be discovered. Thus, we define new prediction features based on the properties of the places visited by users which are able to discriminate potential future links among them.

Building on these findings, we describe a supervised learning framework which exploits these prediction features to predict new links among friends-of-friends and place-friends. Our evaluation shows how the inclusion of information about places and related user activity offers high link prediction performance. These results open new directions for real-world link recommendation systems on location-based social networks.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms

Experimentation, Measurement

Keywords

link prediction, location-based services, social networks

KDD'11, August 21–24, 2011, San Diego, California, USA.

1. INTRODUCTION

Location-based online social networks have seen soaring popularity, attracting millions of users [12]. People are increasingly sharing their location with their friends: such *check-ins* can be broadcasted to friends, while messages, tips or other information related to a place can be generated and shared. As many other online social services, location-based networks may greatly benefit from link recommendation, since as users add more and more friends their engagement with the service increases [15].

An inherent characteristic of these social networks is that they may have millions of nodes, but, at the same time, they are often quite sparse, with low density of links among these nodes. As a result, the link prediction space is so huge and highly imbalanced that real approaches merely focus on finding friends in the the 2-hop social neighborhood, i.e., friends-of-friends of a user [8]. Extending prediction efforts to the 3-hop neighborhood, or even further, may likely result in an exponentially larger set of increasingly less likely candidates. As a consequence, the link prediction problem appears so heavily influenced by network distance between users that each social neighborhood should be treated as a separate prediction problem [17].

Nonetheless, in location-based social networks there is an unprecedented source of potential promising candidates for link prediction: the places visited by each user. Data about the venues where users check-in can be exploited to find and predict future connections. Therefore, the question we tackle in this work is: *how do we design a link prediction system which exploits data about user check-ins?*

To investigate the practical feasibility of such approach we have collected extensive longitudinal data about an online location-based service, Gowalla, with hundreds of thousands of users and more than one million different venues. We have acquired complete data about places, users, their friends and their check-ins for 4 consecutive monthly snapshots (Section 2). We analyze the link prediction space by investigating how new friendship connections are created over time: we discover that *about 30% of all new links appear among users that check-in at the same places.* Thus, these "place-friends" represent disconnected users that can become direct connections (Section 3).

We argue that effective link prediction on location-based services can greatly benefit from focusing only on the friends-of-friends and on the place-friends of a user. This design choice makes the prediction space about 15 times smaller than the entire set of candidates and, yet, it covers about 66% of new social ties. In addition, this reduced prediction

set offers a better balance between the number of new links and its total size. As a result, practical implementation of link prediction systems can become more feasible, since for each user only a much smaller set of potential friends has to be explored to compute predictions.

The challenge is then how to exploit the information given by the check-ins of two users, who do not share any friends but who visit the same places, to predict whether they will become direct connections. In fact, activity and interaction revolving around physical places can result in social ties emerging among individuals and correlated to the properties of the place itself, as the sociological "focus theory" suggests [9]. Hence, we define prediction features which quantify users that are likely to become friends considering the places they visit and the properties of these places. Our prediction features are based on user check-ins and on the concept of "place entropy" [5], which is used to discriminate venues that are more or less likely to foster social connections (Section 4).

We finally describe how such prediction features, combined with other measures, can be used in a supervised learning framework to predict future links in a realistic deployment (Section 5) [17]. Our evaluation shows the effectiveness of our design choices, with AUC values of up to 0.96 when predicting links between users who visits the same places but have no friends in common (Section 6). Our approach can be adopted in any scenario where users of a social service disclose data about their visits to places. We conclude the paper with an overview of related results (Section 7) and with a discussion of future works (Section 8).

2. DATASET

In this section we briefly describe Gowalla and the collection procedure we used to acquire our dataset, presenting some of its basic properties.

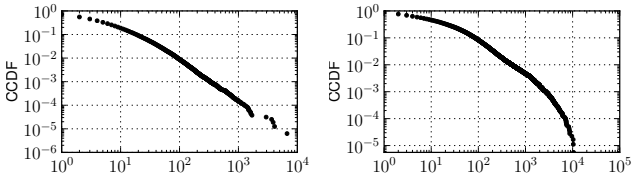
2.1 Data collection

Gowalla is a location-based social networking service created in 2008 that allows users to add friends and share their location with them. The friendship relationship is mutual, requiring each user to accept friendship requests to allow location sharing.

We have downloaded four monthly snapshots of Gowalla data between May and August 2010: we were able to exhaustively query all user accounts, downloading information regarding their profiles, their friends and their past check-ins. We also gathered the geographic location of each place. This dataset represents a sequence of *complete* snapshots of a large-scale online service: this will allow us to study how links are created over time and to evaluate how a link prediction system would perform in a real scenario.

t	Users	Active users	Places	Check-ins
1	252,020	148,234	958,823	7,475,401
2	291,812	168,925	1,104,771	9,073,157
3	325,025	189,512	1,226,847	10,537,516
4	382,750	216,734	1,421,262	12,846,151

Table 1: Properties of our Gowalla dataset across the different temporal snapshots: total number of registered users and active users, total number of different places, total number of check-ins.



(a) Friends per user (b) Places per user

Figure 1: Complementary Cumulative Distribution (CCDF) of the number of friends (a) and of the number of places (b) per user for the last snapshot of the dataset (Month 4). The probability distributions do not change significantly across different snaphots.

t	N	K	G_C	$\langle k \rangle$
1	109,045	476,409	102,951 (94.4%)	8.73
2	124,190	559,901	117,868 (94.7%)	9.01
3	138,387	630,045	131,711 (95.1%)	9.10
4	159,391	736,778	152,011 (95.3%)	9.24

Table 2: Properties of the social graphs at each snapshot: number of nodes N and edges K, number of nodes G_C (and their proportion) in the giant connected component and average node degree $\langle k \rangle$.

2.2 Dataset properties

In the 4 consecutive monthly snapshots Gowalla increased its total number of registered users from about 250 thousands to about 380 thousands, as shown in Table 1, constantly exhibiting about 56% of active users, that is users with at least one friend or one check-in.

As reported in Table 2, each snapshot of our dataset results in a social graph with a subset of the active users: each graph exhibits a large giant connected component, always containing more than 94% of all the nodes. The average number of friends per user grows from 8.73 to 9.24: moreover, as described in Figure 1(a), the degree distribution shows a heavy tail, with only 1% of users having more than 100 friends. Overall the social network is sparse, making link prediction challenging because of the scarcity of social ties. User check-in activity also presents a heavy-tailed distribution: 90% of users with check-ins have visited less than 100 different venues, as detailed by Figure 1(b). Even though users might visit only few places, users who visit the same places are still more likely to become friends than what would be expected on average, as we will see later.

Finally, we note that while many users might have social connections and no check-ins, there are also many accounts with check-ins but no friends at all. On average, *only 57% of active users have both some friends and some check-ins, while 26% have no friends and 17% have no check-ins.* This partition is approximately constant across our temporal snapshots of Gowalla.

3. PREDICTION SPACE ANALYSIS

In this section we study how new friendship connections are created by Gowalla users, exploring how the prediction space can be divided to improve link prediction performance. We will introduce the concept of *place-friends* and show how the search for new social ties can be greatly simplified.

t	U_t	E_t^{NEW}	S_t^{NEW}	P_t^{NEW}	$S_t^{NEW} \cap P_t^{NEW}$	$S_t^{NEW} \cup P_t^{NEW}$
1	148,234	43,812 (100.00%)	24,174 (56.41%)	13,150 (30.01%)	7,677 (17.52%)	30,187 (68.90%)
2	168,925	40,643 (100.00%)	21,118 (51.96%)	12,572 (30.93%)	7,131 (17.54%)	26,559 (65.35%)
3	189,512	58,238 (100.00%)	30,581 (51.51%)	20,107 (34.52%)	10,935 (18.78%)	39,753 (68.26%)

Table 3: Link formation: for each monthly network snapshot we report the total number of active users U_t, the total number of new links appearing among them in the next snapshot E_t^{NEW} and the breakdown of this quantity among new links appearing among friends-of-friends S_t^{NEW} and among place-friends P_t^{NEW}, including the intersection and union of these two latter sets. Percentages are computed with respect to the total number of new links.

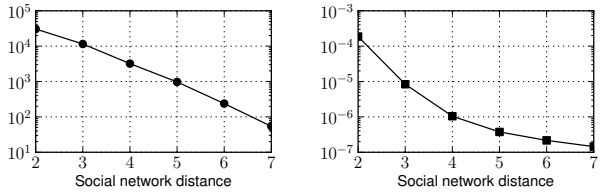
3.1 Definition and notation

Formally, we represent each snapshot of our dataset as an undirected graph $G_t = (V_t, E_t)$ for $t = 1, 2, 3, 4$, where t indicates the different snapshots in time of the dataset. The set of nodes $V_t = \{u_1, u_2, \ldots, u_{N_t}\}$ is composed of N_t users and the set of edges E_t is composed of pairs of users that are present in each other's friend lists in snapshot t. We define Γ_i^t as the set of users connected to user u_i in graph G_t, so that $k_i^t = |\Gamma_i^t|$ is the number of friends of u_i in snapshot t. In addition, there are L_t different places $M_t = \{m_1, m_2, \ldots, m_{L_t}\}$ where users have checked-in at and c_{ij}^t represents the number of check-ins that user u_i has ever done at place m_j until time t, with the number of check-ins of a user in a place able only to increase with time. All the check-ins of user u_i until time t can also be represented as a vector $\vec{c_i^t} = (c_{i1}^t, c_{i2}^t, \ldots c_{iL_t}^t)$. Then, Φ_j^t is the set of all users who have checked-in at place m_j and Θ_i^t is the set of all places where user u_i has checked-in at, both until snapshot t. Finally, $A_t = \bigcup_{j=1}^{L_t} \Phi_j^t$ is the set of all users with at least one check-in at snapshot t, while $U_t = V_t \cup A_t$ is the set of all users present at snapshot t with at least one friend or one check-in.

3.2 Dividing prediction space

Users do not add friendship connections at random with all other users but, instead, tend to prefer other users that are "close" to them, either in social sense or along other dimensions such as geographic proximity or topic interest [15, 1, 6, 19]. For instance, many links do appear between individuals at closer social distance from each other, with the 2-hop neighborhood of single nodes being the largest source of new ties [17]. This seems to hold also in Gowalla: as shown in Figure 2(a), the number of new links appearing between users which are d hops away exponentially decreases with d. Moreover, the likelihood that a couple of users at network distance d will have a link in the next snapshot of our dataset decreases sharply with d, as described by Figure 2(b): the probability that two users with at least one friend in common, thus being at distance $d = 2$, will become friends is above 10^{-4}, but this value quickly tumbles down below 10^{-5} and to 10^{-6} at distance $d = 3$ and $d = 4$ respectively. Hence, pairs of users at larger distances give a weaker contribution to link formation, both in terms of absolute number of new links and likelihood of a new social tie [14].

Nonetheless, in a location-based social network the social dimension is not the only one to be exploited and investigated. Instead, in our context there is an additional source of information about social ties: the places where users check-in. In particular, users may add a new connection not because of a shared friend but because of a shared place.



(a) New links (b) Probability of a new link

Figure 2: Number of new links appearing among pairs of nodes at different values of social distance (a) and their relative probability of appearance (b). Pairs of users at closer distance are both generating a larger number of social links and more likely to turn into social links.

In order to quantify how users seek and add new friends, for each snapshot and for each user u_i we define two sets of potential friend couples:

Friends-of-friends

$$S_i^t = \{(u_i, u) : u \in \Big(\bigcup_{u_k \in \Gamma_i^t} \Gamma_k^t \Big) \setminus \Gamma_i^t\}$$

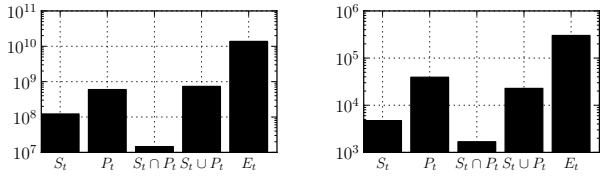
Place-friends

$$P_i^t = \{(u_i, u) : u \in \Big(\bigcup_{m_k \in \Theta_i^t} \Phi_k^t \Big) \setminus \Gamma_i^t\}$$

While friends-of-friends are all those users that share at least one friend without being directly connected, place-friends are all those users that have checked-in at least in one common place but are not connected to each other. These two sets may not be disjoint for a given user u_i. Finally, we define two sets containing all the potential pairs of nodes that are either friends-of-friends or place-friends in a given snapshot: $S_t = \bigcup_{u_i} S_i^t$ and $P_t = \bigcup_{u_i} P_i^t$.

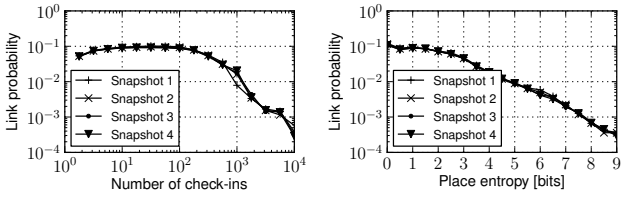
The monthly snapshots of our dataset make it possible to quantify how many new social links appear within these two sets. For every network snapshot $G_t = (V_t, E_t)$ we define $E_t^{NEW} = E_{t+1} \cap ((U_t \times U_t) \setminus E_t)$ as the set of all new links appearing in the next network snapshot $t+1$ among all users already present at snapshot t. In Table 3 new links appearing between temporal snapshots are classified according to their origin: $S_t^{NEW} = E_t^{NEW} \cap S_t$ and $P_t^{NEW} = E_t^{NEW} \cap P_t$ are, respectively, the set of new links among friends-of-friends and the set of new links among place-friends.

About two-thirds of all new links appear within $S_t \cup P_t$. In particular, while about 50% of new links appear among friends-of-friends, more than 30% of new friends are added among place-friends that check-in at the same venues. Fi-



(a) Prediction space size (b) Imbalance ratio

Figure 3: Number of potential friends (a) and imbalance ratio (b) for each class of potential new links: for social potential neighbors S_t, for place potential neighbors P_t, for their intersection and union and for the entire set of users E_t. Results averaged over all temporal snapshots.



(a) (b)

Figure 4: Average probability that two users who have checked-in at a place are friends as a function of the number of check-ins in that place (a) and as a function of place entropy (b).

nally, about 13% of new links appear between users without any friends in common but who are place-friends.

3.3 Reducing prediction space

In addition to the absolute number of new links appearing among friends-of-friends and place-friends, it is also important to study how link prediction feasibility can vary across these prediction spaces. In a prediction space there are both couples of users that will become connected and couples that will not: the performance of prediction approaches depends on the total number of these potential couples and on the relative proportion of these two classes. Exhaustive approaches would scale with the total number of potential links, which can become prohibitively large for real-world online social networks with millions of users. Also, the two classes can present an extremely skewed distribution, with new links being greatly outnumbered by couples of users that will never create a social tie. This problem is worsened by the fact that these new links are actually the occurrences of greater interest, yet they are difficult to find.

In Figure 3(a) we report the prediction space size for the friends-of-friends set S_t and the place-friends set P_t, including also their intersection and union, along with the size of the overall prediction space for the entire dataset. While there are more than 11 billions couples of users, there are about 700 million place-friends (P_t) and about 100 million friends-of-friends (S_t), with their intersection reducing the prediction space to about 20 million entities. Thus, by focusing prediction efforts only on place-friends or friend-of-friends the prediction space can be reduced by about 15 times, while still covering two-thirds of all new links.

Then, we study the *imbalance ratio* of a prediction set, which is the ratio between the total number of items and the actual number of new links that will appear within it. Imbalance ratios are key indicators of link prediction systems performance: they express how many real instances should be considered and analyzed, on average, before a prediction can be successfully done. Place-friends and friends-of-friends offer lower imbalance ratios than the overall prediction space, as presented in Figure 3(b): hence, not only they offer a smaller prediction space, but also the likelihood that new links will be found is about 20 times higher than the average.

However, discovering friends between users who check-in at the same places appears challenging. Not all places have the same importance for different users and, thus, not all places are equally likely to foster new social ties among individuals who visit them. The key idea is then to take advantage of the properties of a place to predict new links.

4. PREDICTION FEATURES

In this section we will describe how place properties can be exploited in link prediction systems and we will introduce the prediction features we adopt in our model.

4.1 Place Properties

Places can be characterized by taking into account users check-in: in fact, the average probability that two users who have checked-in at the same place are friends exhibits a decreasing trend as the place has more check-ins, as shown in Figure 4(a). Yet, there is not much difference when a place has less than 100 check-ins.

However, a place where only a small number of users regularly check-in is likely to be a place with a significant importance for them, such as private houses, gyms, offices. Conversely, a place with a similar total number of check-ins but made by several users is likely to be a public place without considerable significance to its visitors, such as touristic places, airports, train stations and so on.

Hence, a more suitable measure of how much a venue promotes social connections among its visitors should take into account both the number of users that check-in and their number of check-ins. A feasible combination is to exploit information theory and define an entropy-based measure to assess the importance of place for social link creation. *Place entropy* has been used in ecology to measure place biodiversity [5]: the underlying assumption is that a uniform distribution of species in a given physical environment is much more diverse than a skewed distribution, where only a few species are overwhelmingly present.

Let C_k^P be the total number of check-ins all users have at place m_k and $q_{ik} = c_{ik}/C_k^P$ the fraction of check-ins that user u_i has at location m_k with respect to the total number of check-ins at place m_k. Then $\{q_{1k}, \ldots, q_{Nk}\}$ is a discrete probability distribution that describes how likely a check-in at m_k was made by a certain user. Thus, we define E_k as the entropy of place m_k:

$$E_k = -\sum_{u_i \in \Phi_k} q_{ik} \log q_{ik} \quad (1)$$

Venues visited by several casual users are less likely to foster the creation of social links between them. Hence, places with higher entropy might result in less social links among

their visitors than venues with lower values. This is confirmed by Figure 4(b): the average probability that two users who have checked-in at the same place are friends decreases as the entropy of the place itself increases. Place entropy seems to offer a strong discriminative power: as we will see, it is a successful indicator of whether a certain place is likely to result in social ties between its visitors.

4.2 Feature definition

Link prediction methods are based on numeric scores computed for pairs of users. These values tend to capture proximity of two users across different dimensions, with the underlying assumption that couples of users that are similar or close are likely to develop a social connection between them.

We will consider *social features*, which can be computed for friends-of-friends, *place features* which can be computed for place-friends, and *global features*, that can be computed for any couple of users even, if they do not share any friend or place. All features are described in Table 4 and discussed in the following paragraphs.

4.2.1 Place features

When two users check-in at the same places they might have many chances to be in contact with each other and, therefore, to create a new connection between them. The two features `common_p` and `overlap_p` denote respectively the number and the fraction of common places between two users, while `w_common_p` takes into account the number of check-ins of both users and `w_overlap_p` is given by the cosine similarity of the two check-in vectors.

Then, we define two features based on the entropy of the places that two users share: `min_ent`, the minimum place entropy across all the shared venues, and `aa_ent`, the sum of the inverse of each place entropy value, a measure inspired by the Adamic-Adar similarity score [1]. Similarly, we define corresponding features considering the number of check-ins, `aa_p` and `min_p`: in this case the relevance of a shared place is higher if it has only a few check-ins.

4.2.2 Social features

Several link prediction features are based on the assumption that two users that share many common neighbors are more likely to create a direct connection. Thus, given two users we define `common_n` as their number of common neighbors and `overlap_n` as their Jaccard coefficient [21]. In addition, `aa_n` is their Adamic-Adar measure based on the degrees of the shared neighbors [1].

4.2.3 Global features

Finally, we define measures that can be adopted for any pair of users, as they are based on their individual properties.

An approach to link prediction is to consider the geographic distance between two users, since geographic proximity is related to higher chances of social connection [16, 2, 19, 4]. We define m_{l_i} as the "home-location" where user u_i has most check-ins: given two users, we compute `geodist` as the geographic distance between their home locations. At the same time, `w_geodist` is the same distance divided by the product of the number of check-ins each user has done in their home location.

Another method to define global features is to consider how many friends users have added or how many places they have visited. We define `pa` as the preferential attachment score of two users [3], whereas `pp`, or *place-product*, is given by the product between the number of places that each user has visited. These two features tend to capture more active users that tend to visit many places or add many friends.

Place features	
`common_p`	$\lvert\Phi_i \cap \Phi_j\rvert$
`overlap_p`	$\frac{\lvert\Phi_i \cap \Phi_j\rvert}{\lvert\Phi_i \cup \Phi_j\rvert}$
`w_common_p`	$\vec{c_i}\vec{c_j}$
`w_overlap_p`	$\vec{c_i}\vec{c_j}/\sqrt{\vec{c_i}^2\vec{c_j}^2}$
`aa_ent`	$\sum_{m_k \in \Phi_i \cap \Phi_j} \frac{1}{E_k}$
`min_ent`	$\min(E_k : m_k \in \Phi_i \cap \Phi_j)$
`aa_p`	$\sum_{m_k \in \Phi_i \cap \Phi_j} \frac{1}{\log C_k^P}$
`min_p`	$\min(C_k^P : m_k \in \Phi_i \cap \Phi_j)$
Social features	
`common_n`	$\lvert\Gamma_i \cap \Gamma_j\rvert$
`overlap_n`	$\frac{\lvert\Gamma_i \cap \Gamma_j\rvert}{\lvert\Gamma_i \cup \Gamma_j\rvert}$
`aa_n`	$\sum_{z \in \Gamma_i \cap \Gamma_j} \frac{1}{\log(\lvert\Gamma_z\rvert)}$
Global features	
`geodist`	$\mathrm{dist}(m_{l_i}, m_{l_j})$
`w_geodist`	$\mathrm{dist}(m_{l_i}, m_{l_j})/c_{il_i}c_{jl_j}$
`pa`	$\lvert\Gamma_i\rvert\lvert\Gamma_j\rvert$
`pp`	$\lvert\Phi_i\rvert\lvert\Phi_j\rvert$

Table 4: List of prediction features.

5. LINK PREDICTION

In this section we describe our link prediction framework. Our proposal builds on two key choices:

- reducing the prediction space by focusing only on friends-of-friends and place-friends;
- exploiting prediction features based on the places visited by users.

We propose a *supervised learning approach* to link prediction, modelling it as a binary classification problem which adopts the prediction features previously described.

5.1 Prediction candidates

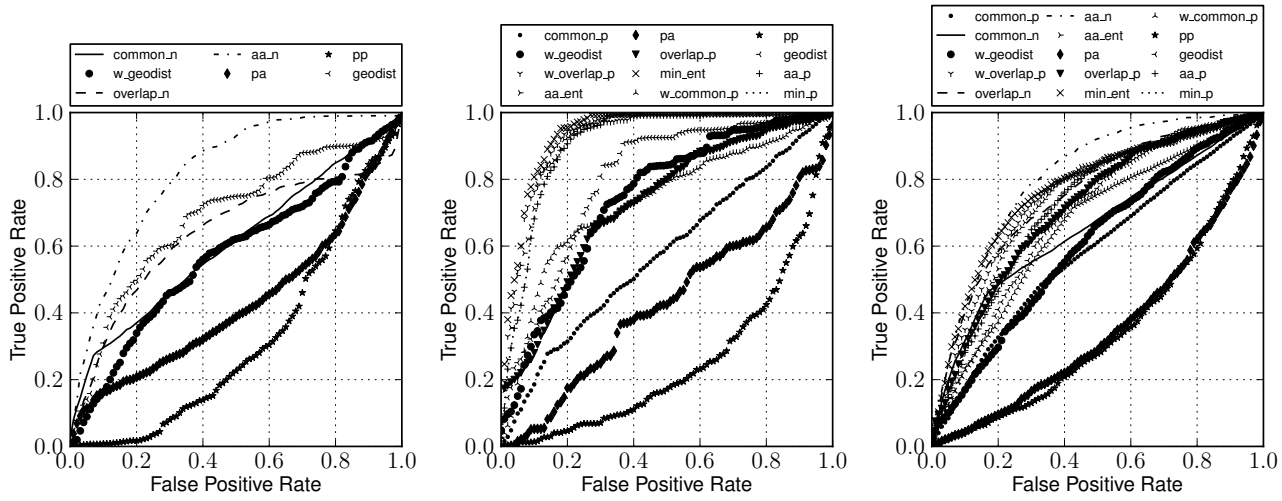
Let us consider a dataset snapshot, with U_t being the set of all users and $G_t = (V_t, E_t)$ the relative social network (as defined in Section 3.1). The link prediction problem can be formulated as: given the dataset snapshot at time t as input, compute and return a set of pairs of users $E_t^{PRED} \subset (U_t \times U_t) \setminus E_t$ that are predicted to appear in E_{t+1}

The entire prediction space $(U_t \times U_t) \setminus E_t$ contains all the potential couples between users that are not yet connected by a link. Exploiting the findings of our previous analysis of this prediction space in Gowalla, we select three disjoint prediction sets:

i) **Social**: links appearing among couples that are friends-of-friends but not place-friends (the set $S_t \setminus P_t$);

ii) **Place**: links appearing among users that are place-friends but not friends-of-friends (the set $P_t \setminus S_t$);

iii) **Place-social**: links appearing among users that are both friends-of-friends and place-friends (the set $S_t \cap P_t$).



(a) Social (b) Place (c) Place-social

Figure 5: ROC curves for individual features used as unsupervised prediction methods on the three different prediction sets.

Our choice is motivated by the fact that combining these three prediction sets results in a set of candidates about 15 times smaller than the entire prediction space while still allowing us to predict two-thirds of new social ties, as discussed in Section 3.3.

5.2 Prediction algorithm

We adopt a supervised learning approach: for every snapshot t, we compute features at time t for couples of disconnected users and we assign a positive label to each couple if they become connected with a link at $t+1$, or a negative label otherwise. Thus, training and test sets are built so that features from a given time interval are mapped to class labels in a future time interval. Hence, given our 4 snapshots, we can create 3 learning sets, each one with labels drawn from the next snapshot.

Classifiers can then be trained to build models and recognize positive and negative items from their features. As motivated by recent results [17], the choice of a supervised learning formulation to address the link prediction problem stems from the heavily skewed distribution of class labels. Unlike unsupervised methods, class distributions are learned by supervised algorithms, allowing a more effective discovery of inter-class boundaries and hence better classification performance.

6. EXPERIMENTAL EVALUATION

We now present the experimental evaluation of our method: this section includes an investigation of the predictive power of each similarity feature and then an analysis about different supervised classifiers which use these features. Our results show how link prediction systems based on our proposal may be feasibly deployed on similar services with high accuracy.

6.1 Evaluation strategy

For each snapshot t and for each prediction set we sample disjoint training and test datasets: these datasets are always sampled to maintain the original unbalanced distribution of positive and negative items in the real data. Finally, for every item we compute all available prediction features: the only limitations are that in the Social prediction set *place features* are not defined and in the Place prediction set *social features* are not defined. All our evaluation tests have been performed with the WEKA framework, which implements several machine learning algorithms, using default parameters (unless otherwise specified) [25].

We adopt Receiver-Operating-Characteristic (ROC) curves as the main tool to evaluate prediction performance [18]. ROC curves describe how the fraction of true positives over all the positive cases changes as a function of the fraction of true negatives over all the negative cases when the decision threshold varies. A ROC plot is a monotonic non-decreasing plot of true positive rate as a function of false positive rate. A random classifier will result, on average, in the curve $y = x$, while better classifiers will result in curves closer to the upper left corner. ROC curves are particularly able to assess classification performance for highly imbalanced datasets, as in our case. The area under the ROC curve (AUC) is often adopted as a scalar measure of the overall performance.

6.2 Individual features evaluation

We study the predictive power of each individual feature: we compute predictive scores for every pair of disconnected users in the test set and then we numerically rank these candidates according to their score. Given a decision threshold, new links are predicted for all the candidates with scores higher (or lower, depending on the directionality) than the threshold. As we vary the decision threshold we get true and false positives, generating a ROC curve: these curves are then presented in Figure 5 for each prediction set.

In the Social prediction space, as shown in Figure 5(a), the best feature is `aa_n`, which dominates the other ones. Interestingly, we observe how the global features `pa` and `pp` perform worse than a random predictor. This denotes how

Algorithm	Set	Precision	Recall	AUC
Model trees	S	0.79	0.28	**0.87**
	P	0.87	0.34	**0.96**
	PS	0.92	0.62	**0.95**
Random forests	S	0.92	0.39	**0.85**
	P	0.95	0.72	**0.87**
	PS	0.98	0.84	**0.92**
J48	S	0.63	0.04	0.62
	P	0.86	0.34	0.90
	PS	0.90	0.64	0.91
Naïve Bayes	S	0.01	0.16	0.74
	P	0.01	0.36	0.92
	PS	0.04	0.22	0.82

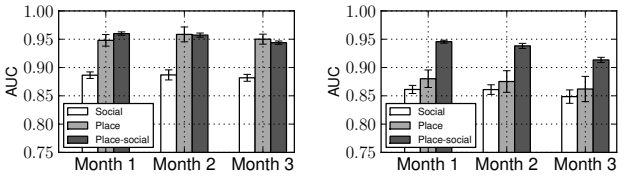
Table 5: Precision and recall on the positive items and overall AUC for different supervised classifiers on the three different prediction sets Social (S), Place (P) and Place-social (PS). Results obtained through 10-fold cross validation and averaged over 20 different random training sets from snapshot $t = 1$.

in the social neighborhood of a given user global indicators are not as useful as measures based on common friends: this may denote how users do not have access to a global view of the network. Instead, global features `geodist` and `w_geodist` perform better, with the former more accurate than the latter. Overall, `aa_n`, `overlap_n` and `geodist` give the best performance, with AUC values between 0.73 and 0.82.

In the Place prediction space, as reported in Figure 5(b), `min_ent`, `w_overlap_p` and `min_p` show the best results, followed by `aa_ent` and `aa_p`. Sharing places with low entropy values or with a few check-ins seems an important indicator of potential friendship, as well as having a large overlap of visited places. These features achieve high AUC values between 0.88 and 0.93. The other features perform slightly worse, with `geodist` doing better than the other ones. Global features `pa` and `pp` show again inverted performance as in the Social case.

Finally, in the Place-social prediction space, as shown in Figure 5(c), all prediction features can be evaluated. As `aa_n` dominates in Social and `min_ent` dominates in Place, they also achieve the best results in this case, with the former having a larger AUC (0.80 against 0.76).

In general, prediction performance is higher in the Place set, while prediction within the other two sets achieves lower AUC values. It is seems easier to predict links among place-friends than among friends-of-friends: this may be due to the fact that more information is available when two users share visited places. However, the prediction space size is much larger in the Place set than in the other two sets, representing an interesting trade-off between prediction effectiveness and search complexity. In essence, the Social set provides good candidates for new links, given its lower imbalance ratio, but then it is difficult to discriminate between them because there is no other information except global features and shared friends. Instead, even if the Place set has higher imbalance ratios, the properties of the places where users check-in provide useful information to discover new friendship connections.



(a) Model trees (b) Random forests

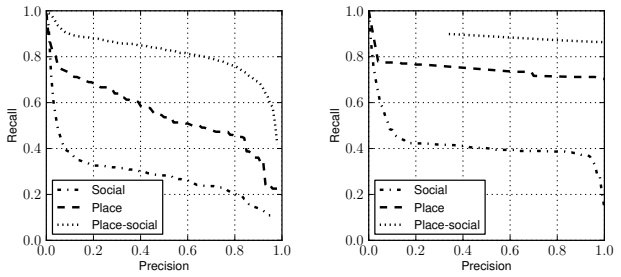
Figure 6: Prediction performance in terms of AUC of model trees (a) and random forests (b) on the three separate Social, Place and Place-social prediction sets, in each temporal snapshots. Results averaged over 20 random datasets, error bars show standard deviation.

6.3 Supervised learning evaluation

We assess whether our prediction features can be combined to characterize a model of link formation across the three prediction sets. Our aim is to achieve at least the same predictive power of the best individual features with a supervised algorithm. We compare the performance of the following classifiers: J48 (equivalent to C4.5[20]), Naïve Bayes, model trees with linear regression on the leaves [10], and random forests (10 trees, 4 random features each) [24]. We run 10-fold cross validation over 20 different training set sampled over each prediction dataset and we consider the AUC value as an overall performance metric [22]. In addition, we also consider two additional metrics computed over positive items: the average *precision*, that is the fraction of positive predictions that are correct, and the average *recall*, that is the fraction of real links that are correctly predicted.

We present our results in Table 5. There is variability across different classifiers: the best performance in terms of AUC are given by random forests and model trees, which are the only two methods that outperform individual features across the three prediction sets (the only exception being random forests underperforming on the Place set). Nonetheless, random forests present higher values of precision and recall than model trees.

As random forests and model trees outperform the other methods, we choose these two classifiers for the next part of this evaluation, where we consider prediction performance across consecutive temporal snapshots of Gowalla. In this case, for every snapshot and for each prediction set we sample disjoint training and test sets of equal size and we compute predictions, averaging results over 20 randomly sampled datasets. As seen in Figures 6, model trees achieve better AUC values on the three prediction sets and across temporal snapshots. Altogether, the two algorithms have lower performance in the Social prediction set, with AUC values between 0.84 and 0.89, whereas Place and Place-social present higher values. Model trees offer slightly better performance than random forests: in particular, the latter algorithm performs worse than individual features on the Place prediction set. A potential explanation for this behavior is that random forests tend to perform poorly when faced with a large heterogeneous set of features, since randomly chosen features are more likely to include less relevant information [11]. This may be the case for the Place set, while this is not the case for Social set, where there are less features, nor for the Place-social set, where there are more fea-



(a) Model trees (b) Random forests

Figure 7: Precision-recall curve for model trees (a) and random forests (b) obtained on the three separate prediction sets, averaged across the three temporal snapshots.

tures but their prediction performance is more homogeneous. However, investigating the precision-recall trade-off offers a different insight on the prediction performance. Given the same level of precision, random forests consistently achieve higher values of recall than model trees, as described in Figure 7. In summary, our prediction framework exhibit high effectiveness with both methods, since they are able to leverage the information contained in our prediction features.

Finally, to understand to which extent different feature classes are contributing to prediction performance we focus only on the Place-social prediction set, where all features are used to build the prediction model, and we test what prediction performance can be achieved by using only one feature class with respect to the full model. As described by Table 6, social features alone provide the worst performance, while both place and global features achieve AUC values closer to the full model. Hence, these two latter classes are mainly contributing to the overall performance, as they exploit information about place check-ins (Place features) and geographic distance between users (Global features). Again, this provides evidence that the choice of including data coming from location-based activity in the prediction model leads to better performance than in purely social-based methods.

6.4 Discussion and implications

Our results are grounded on two main important design choices: focusing link prediction only on a reduced set of candidate pairs of users and exploiting location-based user activity to define successful prediction features. These two simple ideas are able to improve overall performance of link prediction systems: as a consequence, real-world systems can be deployed, making use of predicted links to suggest friends to users and engage them more with the service. In addition, recommending friends among users who check-in in the same places may seem more important in location-based services, since users can directly interact with them when checking-in at these common places.

Our framework enables the prediction of new social ties even for users who do not yet have any friendship connection, provided that they visit and check-in at places. Standard link prediction methods based on social features are of no use in this scenario, since it is impossible to compute prediction features for these isolated users [13]. In some sense, this is a scenario which represents new users of the service: they have signed up, they have checked-in in some places but they are not engaging with other users. Thus, predicting their future links might be extremely important to make them more active participants.

Algorithm	Full model	Social	Place	Global
Model trees	0.95	0.90	0.92	0.92
Random forests	0.93	0.88	0.92	0.92

Table 6: AUC for model trees and random forests on the Place-social prediction set when the full set of prediction features is used and when only a single set of prediction features is used. Results averaged the three snapshots and over 20 different random training and test sets.

Potentially, our proposal to exploit location-based activity to predict new friends could result in improving prediction performance even further by accessing additional information, such as fine-grained temporal information of user activity or direct interaction among users. For instance, users that check-in at the same place and *at the same time* can be much more likely to become friends [4].

7. RELATED WORKS

The link prediction problem in social networks has been under scrutiny for many years. The seminal work by Liben-Nowell and Kleinberg addresses the problem from an algorithmic point of view, investigating how different proximity features can be exploited to predict the occurrence of new ties in a social network [15]. They adopt an unsupervised approach, where scores are computed for all potential candidates and then ranked to obtain the most likely predictions.

More recently, researchers have advocated supervised approaches to link prediction, given the possibility of modelling the task as a binary classification problem. In particular, Lichtenwalter et al. have presented a detailed analysis of challenges in link prediction systems, discussing imbalance problems and proposing to treat prediction separately for different classes of potential friends [17]. While we also adopt a supervised approach, we additionally consider how link prediction can be performed when additional information not arising from social ties is available.

A related approach to find online social ties among mobile users has been presented by Cranshaw et al. [5]: they track a small number of mobile users in the physical world to discover their connections on online social networks. While focusing as well on information-based measures, our approach considers a much larger set of users and studies their activity on a location-based service. Eagle et al. have considered how interactions between people over mobile phones can accurately predict relations among them [7]. Conversely, we do not consider direct interaction nor communication between users to predict social links. A recent work by Crandall et al. [4] shows how temporal and spatial co-occurences between people help to infer social ties among them: while their main goal is to put forward a generative model which explains empirical data, our study has a different aim, that is, designing a link prediction system to be used on real-world location-based services. Furthermore, our work deals with a different type of data: since we exploit check-ins at well-defined venues, we can infer that two individuals vis-

ited exactly the same place without dealing with generic geographic coordinates. As a consequence, our prediction system achieves higher precision while being more feasible for a real-world deployment.

Another thread of research has been addressing the geographic properties of social networks. Liben-Nowell et al. described how the probability of friendship between two individuals can be related to the geographic distance between them [16]. Furthermore, some users on online social networks also exhibit more friendship connections over short geographic distances, with many clusters of friends living nearby [23]. We take advantage of these findings and we explicitly include prediction features based on geographic distance in our prediction framework.

8. CONCLUSION AND FUTURE WORK

In this paper we have described and evaluated a link prediction model based on place properties of a location-based social network. We have studied a large real-world service, Gowalla, finding that the link prediction space can be reduced about 15 times by focusing on place-friends and friends-of-friends only, while still discovering about 66% of all new links. Then, we have described how the properties of the venues visited by users can be used to define prediction features with high predictive power. Building on these findings, we have shown how real link prediction systems can achieve high precision in a prediction space smaller than exhaustive approaches.

Among the future directions of our work we envisage the definition of systems that infer social tie strength from location-based activity and the design of a link recommender system which exploits temporal information about user check-ins.

Acknowledgments

We would like to thank Ilias Leontiadis and Charalampos Rotsos for many useful comments and discussions.

9. REFERENCES

[1] L. A. Adamic and E. Adar. Friends and neighbors on the Web. *Social Networks*, 25(3):211–230, July 2003.

[2] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of WWW '10*, pages 61–70, New York, NY, USA, 2010. ACM.

[3] A. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*, 311(3-4):590–614, August 2002.

[4] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. M. Kleinberg. Inferring social ties from geographic coincidences. *PNAS*, 107(52):22436–22441, 2010.

[5] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the Gap Between Physical Location and Online Social Networks. In *Proceedings of UBICOMP '10*, pages 119–128, New York, NY, USA, 2010. ACM.

[6] N. Eagle and A. S. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, March 2006.

[7] N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *PNAS*, 106(36):15274–15278, September 2009.

[8] Facebook. People You May Know. `http://blog.facebook.com/blog.php?post=15610312130`.

[9] S. L. Feld. The Focused Organization of Social Ties. *The American Journal of Sociology*, 86(5):1015–1035, 1981.

[10] E. Frank, Y. Wang, S. Inglis, G. Holmes, and I. H. Witten. Using model trees for classification. *Machine Learning*, 32:63–76, July 1998.

[11] M. Gashler, C. Giraud-Carrier, and T. Martinez. Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous. In *Proceedings of ICMLA '08*, pages 900–905. IEEE, December 2008.

[12] GigaOM. Foursquare Hits 4 Million Users. `http://gigaom.com/2010/10/21/foursquare-hits-4-million-users/`.

[13] V. Leroy, B. B. Cambazoglu, and F. Bonchi. Cold start link prediction. In *Proceedings of KDD'10*, pages 393–402, New York, NY, USA, 2010. ACM.

[14] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proceeding of KDD' 08*, pages 462–470, New York, NY, USA, 2008. ACM.

[15] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of CIKM '03*, pages 556–559, New York, NY, USA, 2003. ACM.

[16] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *PNAS*, 102(33):11623–11628, August 2005.

[17] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In *Proceedings of KDD '10*, pages 243–252, New York, NY, USA, 2010. ACM.

[18] F. J. Provost, T. Fawcett, and R. Kohavi. The Case against Accuracy Estimation for Comparing Induction Algorithms. In *Proceedings of ICML '98*, ICML '98, pages 445–453, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[19] D. Quercia and L. Capra. Friendsensing: recommending friends using mobile phones. In *Proceedings of RecSys '09*, RecSys '09, pages 273–276, New York, NY, USA, 2009. ACM.

[20] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[21] G. Salton. *Introduction to Modern Information Retrieval*. McGraw-Hill Companies, September 1983.

[22] S. L. Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1:317–327, 1997.

[23] S. Scellato, C. Mascolo, M. Musolesi, and V. Latora. Distance Matters: Geo-social Metrics for Online Social Networks. In *Proceedings of WOSN' 10*, June 2010.

[24] L. B. Statistics and L. Breiman. Random Forests. In *Machine Learning*, pages 5–32, 2001.

[25] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition, June 2005.