Location Affiliation Networks: Bonding Social and Spatial Information

Konstantinos Pelechrinis and Prashant Krishnamurthy

School of Information Sciences
University of Pittsburgh
{kpele,prashant}@sis.pitt.edu

Abstract. Location-based social networks (LBSNs) have recently attracted a lot of attention due to the number of novel services they can offer. Prior work on analysis of LBSNs has mainly focused on the social part of these systems. Even though it is important to know how different the structure of the social graph of an LBSN is as compared to the friendship-based social networks (SNs), it raises the interesting question of what kinds of linkages exist between locations and friendships. The main problem we are investigating is to identify such connections between the social and the spatial planes of an LBSN. In particular, in this paper we focus on answering the following general question "What are the bonds between the social and spatial information in an LBSN and what are the metrics that can reveal them?" In order to tackle this problem, we employ the idea of *affiliation networks*. Analyzing a dataset from a specific LBSN (Gowalla), we make two main interesting observations; (i) the social network exhibits *signs of homophily* with regards to the "places/venues" visited by the users, and (ii) the "nature" of the visited venues that are common to users is powerful and informative in revealing the social/spatial linkages. We further show that the "entropy" (or diversity) of a venue can be used to better connect spatial information with the existing social relations. The entropy records the diversity of a venue and requires only location history of users (it does not need temporal history). Finally, we provide a simple application of our findings for predicting existing friendship relations based on users' historic spatial information. We show that even with simple unsupervised learning models we can achieve significant improvement in prediction when we consider features that capture the "nature" of the venue as compared to the case where only apparent properties of the location history are used (e.g., number of common visits).

1 Introduction

During the last few years, boosted by advancements in mobile handheld devices (e.g., smartphones), a new class of digital social networks, namely location-based social networks (LBSNs), has emerged. It is now possible to bring into the equation of online social networks (OSNs) another dimension, that of ***location***, due to the significantly improved ability of mobile devices to accurately estimate their position or location. The underlying communities not only have social ties (e.g., friendship) and/or interests in common (e.g., sports), but they are also "connected" with regards to their geographic

P. Flach et al. (Eds.): ECML PKDD 2012, Part II, LNCS 7524, pp. 531–547, 2012.

locations (often mapped into "venues" as described later). In other words, LBSNs bond the online and physical social ties through location information.

This bond can enable a number of novel, convenient, and appealing services making LBSNs popular. People can now track their children's locations. By tracking friends, applications such as better coordination for scheduled meetings can be enabled. Applications can also include exploring new places through a list of venues that are within the proximity of the current location. This list can now be accompanied by tips and recommendations from people/friends that have visited these places. Even simply the number of people that have visited a locale in the past or are present at the moment might be helpful and informative. Other systems can also offer Groupon-like deals, providing additional monetary incentives for someone to adopt their usage. A recent study has also shown that "gaming" aspects of LBSNs form an important motivation for people to start using them [12].

With LBSNs becoming prevalent, it becomes critical to comprehend and discriminate the types of knowledge we can obtain from the bond between locations and social ties. For example, what correlations exist between users' spatial trails and their social behaviors as expressed through their friendships and do the spatial trails provide any information about social ties? Our primary objective in this work is to identify the existing correlations and the metrics that can best capture them. Using the knowledge we obtain from our study we further examine *whether we can use these correlations and metrics to infer social information* **only** *from users' locations*. Going forward this can stimulate our ability to deconstruct the interplay between the social and the spatial information plane and apply it to new applications.

Interactions in an LBSN: An LBSN has two distinct components; a social network and a location log for each member. The social part of the system resembles any other existing online social network, where friendships are declared and people can interact with their friends. What differentiates LBSNs from other OSNs are the type of interactions that are feasible between the members of the network. The main feature of this interaction is location sharing. While the "visible" interactions in a traditional OSN are restricted to the virtual world, we can observe interactions within an LBSN in the physical world as well. This is especially important for our study since it can shed light on patterns that are otherwise difficult to identify.

Location sharing can be realized either through continuous tracking, in the form of a temporal latitude/longitude trajectory (e.g., Loopt - see Figure 1) or via "check-ins", where users announce their presence in a place or venue at their convenience (e.g., Gowalla, Foursquare etc. - see Figure 2). Clearly, the second approach, where location is tagged with semantic information as compared to a flat geographic trajectory, offers a richer set of information, but with coarse location granularity. All major LBSNs follow this latter approach and consequently, in this work we consider systems in which spatial information is created via check-ins. We note here that using "check-in" history can be challenging since fine grained temporal information is absent (e.g., users do not "check-out" etc.).

Hence, we now have two types of information – the social ties between members and check-ins of members of the LBSN. To analyze socio-spatial interactions within an LBSN, we model it as an "affiliation network", where the members are nodes of one

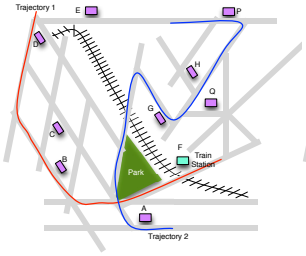


Fig. 1. Trajectory-based LBSN

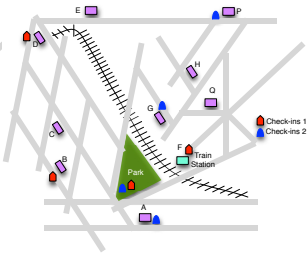


Fig. 2. Check-in based LBSN

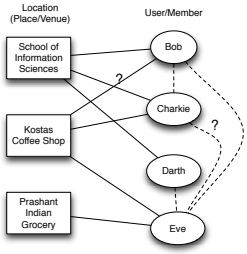


Fig. 3. LBSN as affiliation network

type and venues/places are nodes of the second type (see Figure 3). Using a dataset from Gowalla [2], we analyze how the *number* and *type* of users' common affiliations (as measured through the number of common locales visited by them) are related to the affinities in the underlying social graph. The main contributions of our study can be summarized as follows:

- We identify clear signs of location homophily, that is, members of the LBSN that are friends are more *similar* compared to those that are non-friends. "Similarity" here refers to the **percentage** of visited places that are common between two users (formally defined later).
- While simply the number of common places visited by two users does not provide rich social knowledge, the user similarity as well as the "type" of their common venues is a very descriptive feature.

Using the affiliation network model we are able to define the clustering coefficient (cc) of a venue, which captures the type and diversity of the latter. As we will see later, this cc has a strong correlation with the social relations in the graph; exactly what we are looking for! However, its computation utilizes knowledge from the friendship graph, resulting in the problem of circular reasoning. Hence, we examine other metrics, and in particular we show that the entropy of a venue is very informative and helpful for dealing with our problem.

Finally, we investigate the importance of the different features we consider through simple unsupervised friendship prediction models. In particular, we seek to infer the existing affinity relations using *only* the users' location history. Our evaluations reveal that features that account for the type of a venue, can significantly improve the estimations as compared to features that consider all venues equal.

Scope of Our Study: We would like to emphasize that our work is a study of the interplay between the social and spatial information present in an LBSN. Even though this connection can enable many new applications, such as location prediction, this study is not focused on any specific one of them. Despite the fact that we examine some simple friendship inference models utilizing our findings, our objective in this study is **not** to provide a social affinity classifier but to provide insights into the value of the location information present in an LBSN and its strength for predicting social ties. For instance,

the relation between spatial and social data can have significant implications on users' privacy. Privacy policies that avoid information leakage from one component of the network to the other should be designed and be in place. We believe that this work can stimulate further research and enhance existing – or even enable new – functionalities within an LBSN.

The rest of the paper is organized as follows. Section 2 discusses work related to our study, while Section 3 describes our affiliation network model for an LBSN and the dataset. Section 4 briefly presents the analysis of the social graph of Gowalla. Our study on the relation between users' location information and their social ties is presented in Section 5. Finally, Section 6 presents our friendship inference model, while Section 7 concludes our work.

2 Related Work

There is a set of studies that examine the structural properties of existing LBSNs. In this context, structure refers not only to the properties of the social network graph (as in OSNs) but also to the location component (e.g., physical distance to friends, time and type of check-ins etc.). For instance, Cheng *et al.* [1] use data from Foursquare to examine (i) the spatio-temporal properties of users' check-ins, as well as (ii) their mobility patterns. Similarly, Noulas *et. al.* [16] study the spatio-temporal properties of users activities as captured through the inter-checkin times and the inter-checkin distances. They further identify universal features for human urban mobility [15]. In alignment, Cho *et al.* [2] use cell phone location and LBSN data to understand the laws dictating human mobility. Li and Chen [10] analyze data from Brightkite and after providing the structural properties of the underlying social graph they try to identify correlations between different user's profile features, activity updates, and mobility patterns.

The majority of these studies deal explicitly either with the social part of the system or the location component. Scellato *et al.* [18] try to use information from both components to identify the relation between friendship and geographic distance using data from 3 different LBSNs (Gowalla, Foursquare and Brightkite). They find that the socio-spatial structure of these systems cannot be explained by only geographic factors or only social mechanisms. In addition, there exist a few studies in the literature that examine and analyze the location data present in the system with the goal of revealing undirect, hidden information. Noulas *et al.* [17] obtain a static snapshop of Foursquare in order to analyze the activity in different neighborhoods of London and New York, while Ye *et al.* [22] exploit social and spatial characteristics of LBSNs for location recommendation.

None of the aforementioned works however, study the relation between the location trace of a user and his social relationships. They are all mainly focused on identifying patterns either in the social or in the spatial component of an LBSN. Eagle *et al.* [5] [6], as well as Li *et al.* [11], have developed measures to quantify similarity of users based on their mobility. This similarity can be later used to infer the social structure of the users. They are focused on "co-location instances," that is, situations where two users are at the same place at the same time. However, given the fact that co-locations between people can happen accidentally, especially in urban areas [13], simply accounting for

the number of co-existences can be expected to not be very accurate. Recently, Wang *et al.* [20] using mobile phone data have identified a positive corelation between mobile homophily, network proximity and social tie strength.

The work that is closer to ours is the study by Cranshaw *et al.* [4]. In particular, they introduce the notion of a location's "entropy," which captures its *diversity* with regards to the people visiting it. Using a small scale dataset of location trajectories obtained from 397 users of Locaccino the authors infer co-locations between users. They examine the relation between features such as the intensity and duration of co-locations between people, the diversity of these co-locations, and the users' mobility regularities, with the social structure of these users. The latter is obtained through their Facebook accounts. They apply supervised learning classifiers on a set of 16 features to obtain the structure. Scellato *et al.* [19] took one step further and use location information in order to improve friend recommendations. They are focused on the temporal evolution of the social graph and they utilize a combination of information drawn from both the social and location component to improve friend recommendations. On the contrary, we are focused on a static snapshot of the network and we are looking into relations between the two information planes of the system.

To further differentiate our work, the location information that Cranshaw *et al.* [4] consider includes the *complete trajectory* of users, from which features such as the duration of a co-location can be inferred. Nevertheless, despite the fact that such data include fine-grained spatio-temporal information, note here that such data capture the location of the device and not necessarily that of the user to begin with (a problem identified in [4] as well). On the contrary, in an LBSN such as Gowalla or Foursquare, people check-in to spots declaring their actual presence in a physical location. However, users do not "check-out", making it challenging, if not impossible, to obtain detailed spatio-temporal information (e.g., co-location duration, even actual co-location at all). Yet, as we show later, this information is promising in its ability to tie the spatial and social plane of LBSNs. To summarize, in contrast to the work presented in [4] we are only using the check-in history of users without considering information related to the actual co-locations of users (i.e., temporal information).

3 Location Affiliation Network

In this section we will briefly describe the data set and affiliation network model for the LBSNs used in this paper.

Gowalla Dataset: The dataset consists of 6,442,892 public check-in data performed by 196,591 Gowalla users in 647,923 distinct places, during the period between February 2009 and October 2010. Gowalla users also participate in a friendship network with reciprocal relations, which consists of 950,327 links. The public dataset [2] includes only an ID for the spot of the check-in. We have further crawled the web in order to obtain a mapping between this id and the actual locale (or "spot" in the terminology of Gowalla[1]). Note here that since the acquisition of Gowalla from Facebook, its public

[1] We will use the terms locale, place, venue, spot and affiliation interchangeably.

website is offline. However, we were able to obtain a subset of the required information through the Internet Archive Wayback Machine and Google Cache.

Affiliation Network: Social relations can be formed due to a variety of reasons. For instance, it has been observed that people tend to relate to others with similar characteristics/interests (**homophily**) [9]. When we refer to immutable characteristics it is clear that the main reason behind homophily is the mechanism of **selection** [8]. For instance, people prefer in general to socialize with people of the same nationality. However, when we consider mutable characteristics (e.g., political views) it is not clear whether selection or **social influence** [7] leads to homophily. In the latter case, friendships were first created and then people influenced each other and became similar.

Based on the above, link creation is affected by contextual factors related to the *similarity* between the users. This similarity can refer to characteristics, activities, or behaviors. However, the representation of a social network as a flat affinity graph is not capable of capturing these surrounding contexts. Affiliation networks integrate "focal points" (*foci*) of social interactions with the pure social graph [14]. An affiliation network is essentially a bipartite graph with two sets of nodes, S and F. S is the set of nodes that represents the members/users of the network, while F represents the activities (affiliations or foci) into which users engage. An edge $\{(s, f) : s \in S \wedge f \in F\}$ exists, iff s is participating in focus f. Two users u and v are said to be affiliated if they participate in the same activity f. Hence, the affiliation network becomes the layer on which the actual social network is created. As Watts states, "without any affiliations, the chance that two people will be connected is negligible" [21].

If we further connect members of S based on their social relations, we obtain a *social-affiliation* network (see Figure 3). Using the latter we can analyze the co-evolution of both the social and the affiliation networks. A new friendship might be created due to a common friend (**triadic closure**), or due to a common affiliation (**focal closure**). Furthermore, a new affiliation can be created due to a friend already affiliated with it (**membership closure**). Focal closure is an artifact of the selection process, while membership closure is a type of social influence. In the LBSNs that we consider, the set F consists of the locations/places that people in S can check-in. An affiliation edge is created as long as a user has checked-in a specific spot. For instance, in Figure 3, Bob has checked-into the "School of Information Sciences" and hence there is an affiliation edge that connects him with the corresponding focus.

Before presenting our analysis, we would like to reiterate that data obtained from a system like Gowalla cannot provide fine-grained spatio-temporal information as in [4]. Hence, many of the detailed features used in that study are not available through our dataset. However, even if we do not know whether two friends' affiliations were created simultaneously (i.e., co-location) or with a time lag, their common affiliation is an indicator of a possible relation, and hence a socio-spatial tie. This can be attributed to the fact that both selection and social influence dictate that "friends" will tend to have a number of common affiliations. The difference is that in the former case these affiliations were present when the friendship was formed (and they might have actually caused this social relation to be formed), while in the latter the opposite occurred.

Using the terminology introduced to restate our main objective, we seek to identify patterns/correlations in the social-affiliation network that can reveal ties between the

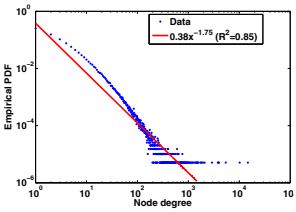


Fig. 4. Node degree distribution

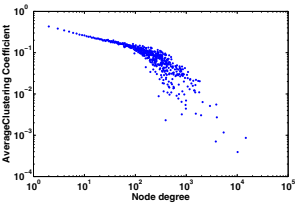


Fig. 5. Clustering coefficient vs node degree

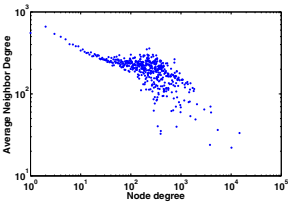


Fig. 6. Average neighbor degree

pure social and pure affiliation network. Note again that when we have a static snapshot of a network, we do not know whether an affiliation or a friendship was created first. However, once again, the actual underlying mechanism that caused the closure between two users and a group is irrelevant and what matters is the existence of a *triangle* that connects users and locales.

4 Social Network Analysis

In this section, we will briefly present our analysis of the social (friendship) graph of Gowalla. There exist similar efforts in the literature for other online social networks and hence this is not the main focus of our study. However, we are presenting these results for completeness.

Degree Distribution: First, we examine the degree distribution of the network. We compute the empirical probability density function of a user's degree (Figure 4). The degree distribution of Gowalla users follows a power law. This has been found to be true for other social networks as well [21], and implies that the majority of the users have very few friends (even none), while very few users have many friends. Formally put, the probability of a node u having a degree of x obeys the following rule:

$$Pr\{deg_u = x\} \propto \frac{1}{x^\alpha} \tag{1}$$

In Figure 4 we have also fit a power law curve, with an exponent of $\alpha = 1.75$. The high R^2 value ($R^2 = 0.85$) implies a good fit, which further supports the validity of the underlying power law. The average node degree is also computed to be 9.66.

Clustering Coefficient: Clustering coefficient (cc for short) is tightly related to the notion of triadic closure. In particular, the clustering coefficient of Bob is an indicator of how many triangles he participates in. Given that the clustering coefficient of Bob is the ratio between the pair of his friends that are friends with each other, over all the possible pairs between them, it is evident that it needs to be presented as a function of the node degree. Figure 5 presents the (average) clustering coefficient of a user with respect to his degree. As we can see, Gowalla users in general exhibit high coefficients, with the average clustering coefficient being equal to 0.237. This means that on average there is a 23.7% probability that two randomly selected friends of Bob will also be friends. This

fairly high clustering coefficient, in conjunction with the small average path length, are strong indications that the social component of Gowalla is a *small world network*.

Average Neighbor Degree: The average neighbor degree $d(k)$ is a summary statistic of the joint degree distribution. It is simply the average neighbor degree of the (average) k-degree node. Figure 6 depicts $d(k)$. As we can see there is no *preference* of users to connect to peers with dissimilar or similar degrees. This can be also captured from the assortativity coefficient of the graph which is close to 0 (-0.029). The slight negative value indicates a very small degree of disassortativity; there are slightly more links connecting nodes of dissimilar degrees.

5 The Richness of Location Information

In this section we will analyze the structure of the spatial component of the LBSN. Our goal is to identify existing correlations, if any, between location information or spatial behavior (represented by the affiliations or checkins at various venues) and the social structure of the network. We are mainly interested in both direct and indirect information derived from location history. For instance, the number of common venues visited by users belongs to the first category. However, information related to the *nature* of the venue is not directly observable from the trails, but it can be inferred.

Location-Based User Similarity: As previously mentioned, homophily is a phenomenon that is very often observed in social networks. For instance, empirical studies have shown that teenagers tend to create friendships with other teenagers with similar scholastic performance and delinquent behavior (e.g., drug use) [8]. In another study, Christakis and Fowler [3] found that social relationships exhibit signs of homophily with regards to the obesity level, in a social network consisting of approximately 12,000 people. Regardless, of the reasons behind homophily, awareness of its existence can help towards revealing possible social links by observations of people's characteristics and/or behaviors and vice versa. To the best of our knowledge there is no study to date that examines homophily related to the locations visited by people. In what follows, we take a first approach to this problem. Our analysis indicates that there are ***signs*** of homophily with regards to the spatial behavior of the users. However, we would like to particularly emphasize that we do not claim to have completely answered this question. Identifying homophily in a social network is an extremely challenging task, which would require the study of longitudinal data, possibly from different networks, on a much larger scale. We hope though, that our work will encourage further research on this topic, which becomes increasingly important nowadays more than ever, with the prevalence of mobile devices with positioning capabilities and the availability of huge volumes of spatial data.

Let us define L_c, to be the set of venues that user c has checked-in. Then we define the similarity $s(u, v)$ between u and v (who have each visited at least one venue) as the following ratio:

$$s(u,v) = \frac{|L_u \cap L_v|}{|L_u \cup L_v|} \tag{2}$$

The numerator is the number of common places visited by the two users, while the denominator is the number of places visited by at least one of them. The above ratio is

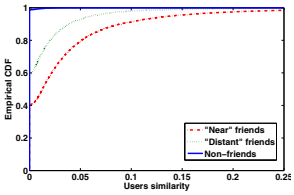


Fig. 7. Empirical CDF for user similarity

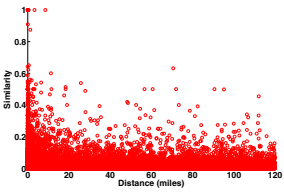


Fig. 8. Near by friends' similarity vs distance

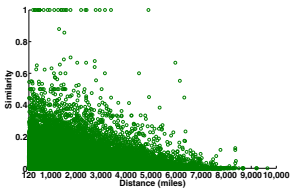


Fig. 9. Distant friends' similarity vs distance

the Jaccard similarity coefficient. We have calculated this ratio for pairs of users that are friends and pairs of users that are not friends. We have also further distinguished the pairs of users as being in geographic proximity or not, based on their "home" locations. We have set up a threshold of 120 miles for defining pairs that are "nearby" or "distant".

Figure 7 presents the similarity for three classes of pairs; nearby pairs of friends, distant pairs of friends and nearby pairs of non-friends. Clearly, friends that reside in geographic proximity to each other, have the highest similarity scores. Approximately, 35% of them have coefficients larger than 5%, which means that 5% of the places they have visited are common. This number might seem small, but it is actually fairly large if we think of the number of places we visit every day. The importance of this value becomes even more clear when we see the similarity index for nearby pairs of non-friends, which is practically 0 even though they are in geographic proximity! Note here that, even friends that are far away, exhibit similarity much higher than nearby pairs of users that are not friends. This is an important result since it implies evidence of homophily in the network with regards to the places visited. Users that are friends will visit the same spots, even if their home locations are far apart. Users that are not friends, even if they are in proximity (e.g., in the same city) are unlikely to visit the same places.

Note here that in the definition of users similarity (Equation 2), we have not considered any temporal information. We consider all common venues that have been visited by two users, regardless of whether they visited them at the same time or not. The reason for this, is that people can be *similar* in ways that do not dictate co-location. For instance, if the selection process is responsible for the high similarity values, people with the same affiliations (captured from the places they visit) will tend to create friendships. On the other hand, if social influence is responsible for the high similarity coefficient, people will tend to visit places that they have heard from their friends (however, not necessarily with them). Hence, the Jaccard similarity index can be quite helpful in bonding social and location information, even without the fine grained temporal information used in previous works. The importance of this finding is that it indicates that the characteristics of location information can be substantially different between friends and non-friends.

Next, Figure 8 shows the similarity values for nearby friends as a function of distance between home locations. As we can see, distance does not appear to have any effect on the similarity for these users. A slight decrease of the (average) coefficient can be observed, but it is not significant. However, distance appears to be critical for friends that live far apart (as one might have expected). As we see in Figure 9, after some

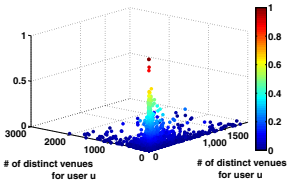


Fig. 10. Similarity as a function of the users distinct affiliations

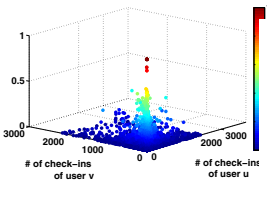


Fig. 11. Similarity as a function of the users checkins

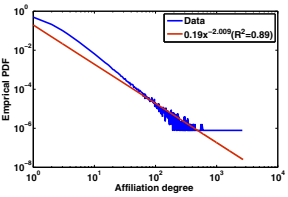


Fig. 12. Affiliation degree distribution

distance (approximately 2,500 miles) the similarity values are drastically reduced. Bob will have less opportunities to "follow" the trails of his friend Alice if she lives far away. Due to space limitations, we do not plot the similarity indexes for the pairs of non-friends, but as one can see from the cumulative distribution function, the majority of these points lay slightly above or on the $y = 0$ line irrespective of the distance.

Figures 10 and 11 present the similarity of two near friends as a function of the number of their distinct affiliations and their check-in counts respectively. Even though our data consist of a static snapshot, this figure can be seen as an "emulation" of the temporal evolution of the similarity value of two nearby friends. Higher levels of activity represent later points in time, when users have been using the system for longer periods and thus, have more affiliations and check-ins. Further, the similarity scores take their maximum values for pairs of users with low levels of activity (i.e., small number of affiliations and check-ins). Based on the above "temporal emulation" this corresponds to early stages of system adoption. This behavior can be attributed to the fact that the denominator of Equation 2 increases faster as compared to the enumerator. One possible reason that can cause this is as follows. Consider Bob and his friend Alice. Bob will hear from Alice about a few places and he will tend to visit some of them, increasing the numerator of $s(Bob, Alice)$. However, he will hear about other spots from his friends Jack and Jill (who might have no relation with Alice). Hence, he might be tempted to visit some of these spots as well, increasing the denominator faster and overall reducing the Jaccard index as his (and Alice's) level of activity increases. Therefore, even friends might exhibit low(er) similarity scores after some time, and for this reason the absolute number of common foci might be a more robust metric over longer time spans. Later in Section 6 we will use $|L_u \cap L_v|$ as the feature of our baseline social link prediction. The similarity of a pair of users, balances the above quantity, by considering the activity of both users. Such a balancing, in essence, captures the diversity of the two users; the larger the denominator, the more places they visit (more diverse user pair). As we will see in our evaluations, this balancing can provide better connection between social and spatial information. We note here that similarity values of non-friends are small regardless their level of activity (omitted due to space limitations).

Focal Closure: In an affiliation network, the foci are also nodes of the network. Hence, we can define metrics such as the degree distribution and the clustering coefficient for venues. The degree of an affiliation-node l (i.e., a venue) is the number of distinct users that have visited it. In other words, it is the number of user-affiliation links whose one endpoint is l. Figure 12 depicts the affiliation degree distribution, which as we can see

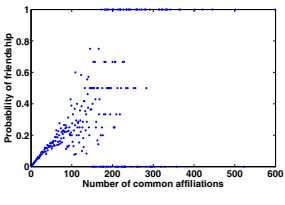


Fig. 13. No clear correlation exists between # of common foci and friendship probability

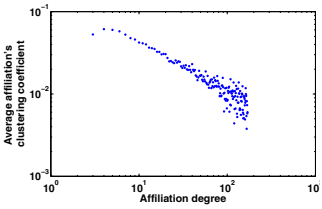


Fig. 14. Venues with lower degree tend to have lower clustering coefficient as well

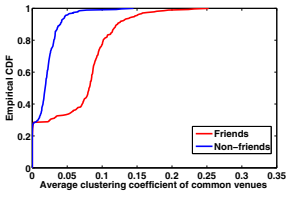


Fig. 15. Common venues of friends have larger clustering coefficient as compared to non-friends

Table 1. Top and bottom 5 venues based on their degree

Top-5 spots	SFO airport	Stockholm Central Station	AUS airport	DFW airport	LAX airport
Bottom-5 spots	"Room"	Farmer's Market	Gas station	Apparel store	Convenient store

follows a power law as well, with exponent $\alpha = 2.01$ ($R^2 = 0.89$). There are a few places with many visitors, while there are many venues with few visitors. The average focus degree is 3.11. Table 1 has the top and bottom 5 venues with regards to their degree. Note here that the bottom 5 venues were randomly selected since there are many venues (almost half) with degree of 1. The top spots are all major transportation hubs (airports or train stations), while the less popular places are more *localized*/personal venues (e.g., "room", which most probably refers to a home, office etc.). The top degree spots are expected to increase the number of common affiliations for many users; the higher the degree of a locale more user pairs will exhibit an increased number of common foci. However, as one can imagine, common affiliations such as a big airport happen rather randomly than due to actual similarity. On the contrary, if Bob and Jack have a local food joint as common affiliation it is highly possible that this is due to their similarity (e.g., they have the same gastronomical preferences).

If our above claim does not hold and all affiliations are *equal* one should expect that the more common foci two people have, the more probable it is for them to be friends. However, our data indicate that this is not the case in an LBSN. Figure 13 presents the friendship probability between a pair of users with respect to the number of common venues visited by them. As we can see there is no clear connection between the two quantities. For small values of common venues there seems to be a linear relationship, but as the number of common affiliations increases, there is little (if any) correlation. In particular, for a number of common venues smaller than 100, the correlation coefficient is fairly high (0.61). However, for larger number of common venues, this coefficients drops to just 0.2. This further supports our previous claim, that the actual affiliation, rather than just the number of the common affiliations, plays an important role in predicting the social relations.

In order to further examine the role of the type of a venue on the social relationships we examine the clustering coefficient of a focus. Let us consider venue l which has a degree of $k > 1$. All the possible social links between the users affiliated with l are

$k(k-1)/2$. If n of them exist then the clustering coefficient, $CC(l)$ is defined as:

$$CC(l) = \frac{n}{k(k-1)/2} \tag{3}$$

This clustering coefficient captures the nature of the place in many ways. It expresses how tightly connected are the people that visit this venue. The higher the cc is, the more connected are the people affiliated with it.

Figure 14 shows the clustering coefficient as a function of the venue's degree. As we can see venues with lower degree have a higher average clustering coefficient. This is a sign that venues with lower degree might venture socialization. Delving more into this issue we present in Figure 15 the CDF of the average clustering coefficient of the common venues for user pairs that are friends and those who are not. For friends, the average clustering coefficient of their common affiliations is much higher (mean value is 0.068) as compared to those of non friends (mean value is 0.019). Finally, Figure 16 plots the probability of friendship between two users as a function of the average clustering coefficient of their common spots. As we see there is a clear positive correlation between the two quantities, which is also revealed from the high correlation coefficient between the two variables (calculcated equal to 0.89). The higher the average cc of the common foci, the larger the friendship probability.

Especially, the last result clearly indicates that the actual *nature* of the venue plays an important role to whether affiliated users are related through a friendship or not. On the one hand, places with high clustering coefficient, attract sets of people that are more tightly connected in the social plane. In addition these sets are usually small, if we recall the connection between affiliation cc and affiliation degree. On the other hand, spots with low clustering coefficient attract many people that are not socially related, just because these places have special features (e.g., large hub-airports, train stations, supermarkets etc.). One could arguably compute the average cc of the common affiliation of two people and find the probability of friendship through a simple linear regression model (Figure 16).

However, there is a problem with the above approach. In order to calculate the average cc of a venue, the social relationships need to be known! Hence, the cc does not provide an **independent** socio-spatial information linkage. Therefore, we need to find a feature of the affiliations, that (i) captures the nature of the venue, (ii) does not require the knowledge social relationships in order to be computed and (iii) is correlated with the friendship probability. This feature is the affiliation's entropy as we will describe in what follows.

The entropy of a place: Cranshaw *et al.* [4] were the first to introduce the notion of entropy of a location as a measure of its diversity. If $P_l(u)$ is the fraction of check-ins in affiliation l contributed by user u, then the entropy of l is given by:

$$e(l) = -\sum_{u:u\in S \wedge P_l(u)>0} P_l(u) log(P_l(u)) \tag{4}$$

From Equation 4 we can see that when a place is visited by many people in equal (and thus, small) proportions, its entropy will be high. In other words, high entropy corresponds to places like airports that exhibit large diversity. On the other hand, when

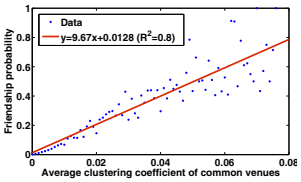


Fig. 16. Positive correlation between avg cc of common venues and friendship probability

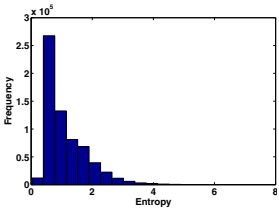


Fig. 17. The majority of the venus exhibit low entropy

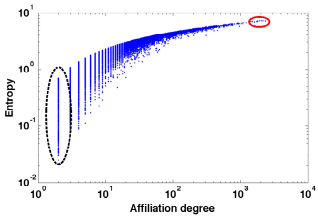


Fig. 18. The entropy of a spot increases with its degree

the mass of $P_l(u)$ is concentrated only to a few people, the diversity in this location is small and so is the entropy.

Figure 17 depicts a histogram of the entropy values for all the venues in our dataset. As we see most of the venues have small entropy values, while there are some that exhibit high entropy values. It is interesting to see that there is an increasing trend of the entropy of a place with its degree (Figure 18). Furthermore, the top-5 degree places are also the top-5 entropy places (with different ranking) as it can be seen in the red solid ellipse. A number of (the many) bottom-5 degree places are still bottom-5 entropy places. However, if we notice more carefully in the dashed, black ellipse in Figure 18, some venues with the lowest degree, do not exhibit the lowest entropy (although still smaller than 1).

Previously, we observed that there is a positive correlation between the average clustering coefficient of the common venues of two users and their friendship probability. To examine whether entropy is a good candidate for a similar correlation, we first examine its relation to cc. Figure 19 depicts the entropy of a venue as a function of its clustering coefficient. As we can see, the entropy tends to be lower as the clustering coefficient increases. High entropy translates to more random co-visits to the venue, and therefore a lower clustering coefficient. Hence, there appears to be a negative relation between these two measures (we expect a similar negative relation between the average entropy of common venues and friendship probability).

Since entropy appears to have similar characteristics with the affiliation clustering coefficient we want to further examine its ability to bond affiliation and social information. In Figure 13 we identified that the number of common affiliations is not very useful in terms of inferring the existing social relations especially when the number of common affiliations is growing (e.g., > 100). We seek to further examine if we can obtain any additional knowledge by utilizing the information about the entropy. Using the same data, we consider pairs of users that have the same number of common foci. We divide them into two categories, friends and non-friends. For each one of these categories we compute the average entropy of the common spots visited and we plot the results in Figure 20. It is clear now that the average entropy of the common affiliations for the case of pairs of friends is indeed lower compared to the case of non-friends and appears to be a good candidate for bridging the social and spatial components of an LBSN.

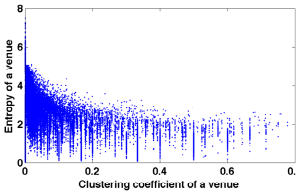


Fig. 19. Venues with higher clustering coefficient tend to have lower entropy

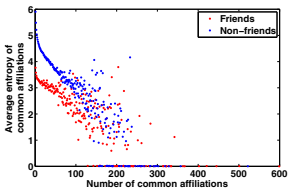


Fig. 20. The common affiliations of friends exhibit lower (average) entropy

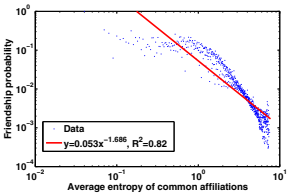


Fig. 21. Negative correlation between the avg entropy of common foci and friendship probability

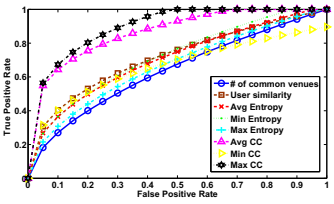
Next, we compute the average friendship probability between two users as a function of the average entropy of their common affiliations. The results presented in Figure 21 are promising. There appears to be a significant (negative) correlation between these two variables (correlation coefficient is equal to -0.7). Their relation follows a power law with exponent $\alpha = -1.69$, as compared to the linear relation between the average cc of the common venues and the friendship probability (Figure 16). This last result, further supports our above argument that the entropy of a place can be used to tie the two information planes and drive applications such as revealing social affinities from location histories. We will examine the latter in the following section.

6 Revealing Friendships

In the previous section, we have examined the user similarity and various venue-related metrics and their correlation with the users' social relations. To summarize, the feature that appears to be able to capture the best the interplay between the social and spatial components of an LBSN is the cc. However, as explained in Section 5, its strong correlation might be illusive, since its calculation explicitly utilizes the social relationships. We further found that the entropy of the common places visited by two users appears to be correlated with their probability of friendship as well and that friends tend to have higher similarity scores. Our analysis therefore implies that similar metrics can be used to make an educated judgment with regards to the social relationship between two randomly selected users whose spatial behaviors are known in terms of the common venues and their entropy. Alternatively, by utilizing a mix of social and partial spatial location it might be possible to estimate future visits of users. The list of possible applications realized through the bonds between social and spatial information in an LBSN is long and not the focus of our study.

In this section, we want to examine the importance of the metrics we considered for estimating the existing users' affinity relations. In other words, considering the graph in Figure 3 and assuming we are only aware of the solid edges, can we estimate the dashed ones? To reiterate, our goal is not to be able to provide a full fledged clustering algorithm on the affinities/ties of the social graph; we only want to examine the strength of the explored metrics in estimating social relations.

Metric	AUC
# common venues	0.63
User similarity	0.73
Avg entropy	0.71
Min entropy	0.76
Max entropy	0.66
Avg cc	0.88
Min cc	0.65
Max cc	0.94



Metric	Precision	Recall
# common venues	0.11	0.016
User similarity	0.28	0.25
Avg entropy	0.42	0.59
Min entropy	0.44	0.63
Max entropy	0.37	0.46
Avg cc	0.89	0.23
Min cc	0.85	0.3
Max cc	0.85	0.3
Full entropy	0.42	0.63
Full cc	0.83	0.33
All features	0.52	0.6

Fig. 22. The predictive power of each considered feature using simple unsupervised learning algorithms: (i) threshold-based (Left table and figure in the middle) and (ii) k-means (Right table). Full entropy (cc) refers to using all the entropy (cc) related features.

We first consider a simple unsupervised, threshold-based, inference model. In particular, for every pair of users, we compute the following metrics, (i) number of common venues (our baseline), (ii) user similarity, (iii) average/min/max entropy of common venues, and (iv) average/min/max cc of common venues. Then, based on a threshold comparison we classify the pair as being friends or not and we obtain the (fitted) ROC curves for the positive instances presented in Figure 22. We focus on the positive (friends) instances, since there is a strong unbalanced distribution of the friends/non-friends instances in the network. Hence even a simple classifier that states every pair as non-friends, would exhibit a very good overall performance, but it would perform very poorly in the classification of the instances of friends. The table on the left also provides the area under the curve (AUC); the larger the AUC, the better is the quality we have in our assessments (lower false positives and larger true positives). As we expected, the average and min cc provide the best performance, while the entropy metrics together with the user similarity come right after, performing better than the baseline of simply the number of common venues. Establishments that are less diverse in terms of people that socialize there tend to be better indicators of bonds, and this information can be used to accurately infer social relations. Furthermore, balancing the number of common venues between two users with their activity improves the prediction, since it accounts for the user pair's diversity as explained earlier.

We further examine an unsupervised clustering algorithm, that operates on the same set of features. We use a simple k-means algorithm and compute the precision and recall on the positive instances. Briefly, precision is the fraction of friendship predictions that are correct, while recall is the fraction of actual friendships that the algorithm was able to identify. The results are presented in the right table at Figure 22 and as we can see again, the features that consider the type of the venue can significantly improve our inference capability. For example, by using the average entropy of the common venues visited by two people we are able to recover 63% of the actual friendships, while from all our friendship predictions, 44% of them are correct. The corresponding percentages when using only the number of common affiliations, are only 1.6% and 11%; a significant improvement. Our results clearly indicate that metrics such as the

entropy and the cc of a venue can help towards the improvement of functionalities such as relationships inference, location prediction etc. It is interesting to observe, that the user similarity performs fairly poor in this test. The reason for this is that there are still friends with low similarity, who are clustered together with the non-friends (low recall), while there are a few non-friends who exhibit larger similarity values and are clustered together with the friends. However, since the friend instances are extremely small in number (as compared to the non-friends), this leads to an overall low precision.

Based on the above results, we believe that defining a user similarity metric that accounts for the entropy of the visited venues can further improve the performance. We seek to examine similar approaches as part of our future work.

7 Conclusions

In this paper we model an LBSN as an affiliation network and by analyzing data from a commercial network we identify bonds between the social and spatial information plane of the system. We find that friends exhibit in general much larger similarity with regards to the number of common venues visited, as compared to non-friends. Considering only the number of common venues between two users, is not very helpful for strongly tying the two components of the network. Even though user similarity can provide a better bonding, the diversity of these common venues with regards to people visiting them is more informative and connect these two parts better. This is also supported by the evaluations and results from simple, unsupervised social link classifiers.

In the future, we seek to examine the location homophily issue in greater detail using longitudinal data and various different similarity metrics (e.g., cosine similarity on an appropriately defined feature vector). As aforementioned, we also opt to intergrate information for the venue entropy in the user similarity and examine any performance improvements in bonding social and spatial information of an LBSN.

References

1. Cheng, Z., Caverlee, J., Lee, K., Sui, D.Z.: Exploring millions of footprints in location sharing services. In: ICWSM (2011)
2. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: Friendship and mobility: User movement in location-based social networks. In: ACM KDD, pp. 279–311 (2011)
3. Christakis, N.A., Fowler, J.H.: The spread of obesity in a large social network over 32 years. New England Journal of Medicine (2007)
4. Cranshaw, J., Toch, E., Hong, J., Kittur, A., Sadeh, N.: Bridging the gap between physical location and online social networks. In: UBICOMP (2010)
5. Eagle, N., Pentland, A.: Eigenbehaviors: identifying structure in routine. In: Behavioral Ecology and Sociobiology (2009)
6. Eagle, N., Pentland, A., Lazer, D.: Inferring friendship network structure by using mobile phone data. Proceedings of the National Academy of Sciences (2009)
7. Friedkin, N.: A structural theory of social influence. Cambridge University Press (1998)
8. Kamdel, D.B.: Homophily, selection and socialization in adolescent friendships. American Journal of Sociology (1978)

9. Lazarsfeld, P.F., Merton, R.K.: Friendship as a social process: A substansive and methodological analysis. In: Freedom and Control in Modern Society (1954)
10. Li, N., Chen, G.: Analysis of a location-based social network. In: IEEE CSE (2009)
11. Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., Ma, W.Y.: Mining user similarity based on location history. In: ACM GIS (2008)
12. Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., Zimmerman, J.: Im the mayor of my house: Examining why people use foursquare - a social-driven location sharing application. In: ACM CHI (2011)
13. Miklas, A.G., Gollu, K.K., Chan, K.K.W., Saroiu, S., Gummadi, K.P., de Lara, E.: Exploiting Social Interactions in Mobile Systems. In: Krumm, J., Abowd, G.D., Seneviratne, A., Strang, T. (eds.) UbiComp 2007. LNCS, vol. 4717, pp. 409–428. Springer, Heidelberg (2007)
14. Newmann, M.E.J., Watts, D.J., Strogatz, S.H.: Random graph models of social networks. Proceedings of the National Academy of Science (2002)
15. Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., Mascolo, C.: A tale of many cities: universal patters in human urban mobility. PLoS ONE 7(5), e37027 (2011), doi:10.1371/journal.pone.0037027
16. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: An empirical study of geographic user activity patterns in foursquare. In: ICWSM (poster session) (2011)
17. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In: SMW (2011)
18. Scellato, S., Noulas, A., Lambiotte, R., Mascolo, C.: Socio-spatial properties of online location-based social networks. In: ICWSM (2011)
19. Scellato, S., Noulas, A., Mascolo, C.: Exploiting place features in link prediction on location-based social networks. In: ACM KDD (2011)
20. Wang, D., Pedreschi, D., Song, C., Giannotti, F., Barabasi, A.-L.: Human mobility, social ties, and link prediction. In: ACM KDD (2011)
21. Watts, D.J.: Six degrees: The science of a connected age. W.W. Norton & Company (2003)
22. Ye, M., Yin, P., Lee, W.-C.: Location recommendation for location-based social networks. In: ACM GIS (2010)