Socio-spatial affiliation networks

Konstantinos Pelechrinis*, Prashant Krishnamurthy

University of Pittsburgh, Pittsburgh, PA 15260, USA

ARTICLE INFO

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" 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 or supervised 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).

© 2015 Elsevier B.V. All rights reserved.

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 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 [15].

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

---

* Corresponding author. Tel.: +1 4126249417.
E-mail address: kpele@pitt.edu (K. Pelechrinis).

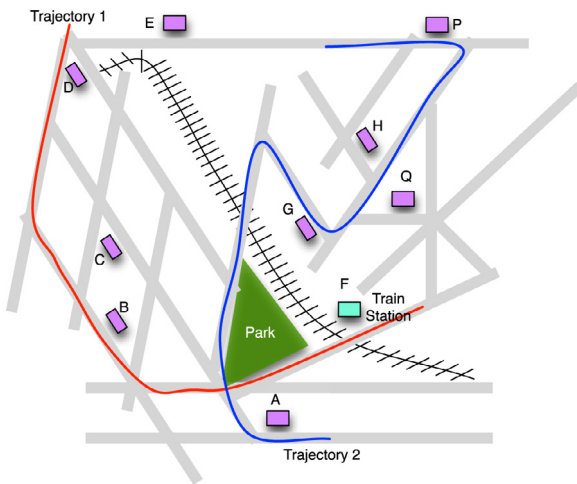**Fig. 1.** Trajectory-based LBSN.

Fig. 2. Check-in based LBSN.

Fig. 3. LBSN as affiliation network.
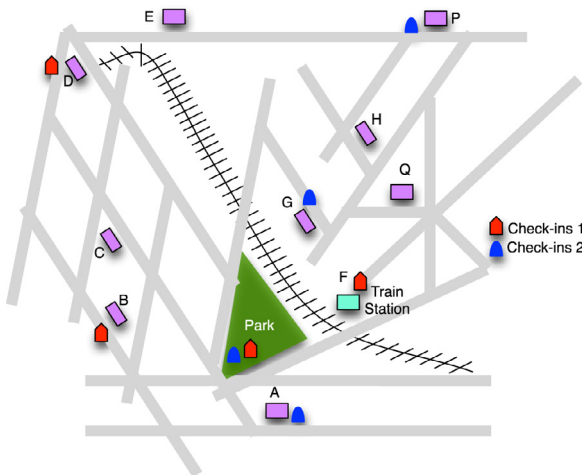
the system resembles any other existing online social network, where friendships are declared and people can interact with their friends. What differentiates LBSNs from other OSNs are the type of interactions that are feasible between the members of the network. The main feature of this interaction is 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 Fig. 1) or via "check-ins", where users announce their presence in a place or venue at their convenience (e.g., Gowalla, Foursquare etc. - see Fig. 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 type and venues/places are nodes of the second type (see Fig. 3). Using a
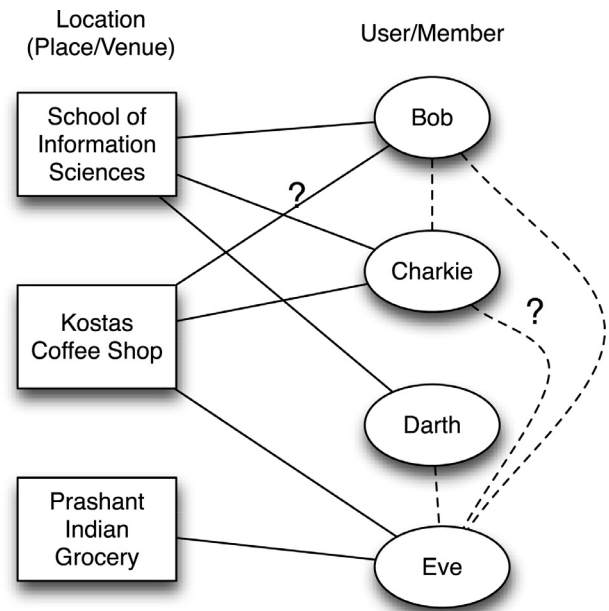
dataset from Gowalla [2], we analyze how the *number* and *type* of users' common affiliations (as measured through the number of common locales visited by them) are related to the affinities in the underlying social graph. The main findings of our study can be summarized as follows:

- 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 (to be 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 can capture the nature of a place in a variety of ways to be elaborated on later. 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 and supervised 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 ability towards 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 describes our affiliation network model for an LBSN and the dataset, while Section 3 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 4. Section 5 presents our friendship inference model, while Section 6 discusses work related to our study. Finally, in Section 7 we discuss some of the limitations of our work, while Section 8 concludes our work.

2. Network model and analysis setup

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

2.1. Gowalla dataset

The dataset consists of 6,442,892 public check-in data performed by 196,591 Gowalla users in 1,280,969 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 website is offline. However, we were able to obtain a subset of the required information through the Internet Archive Wayback Machine and Google Cache.

User's home location: The dataset does not include explicit home location information for the users. In order to infer the home locations of the users, which will be needed for part of our analysis, we apply a density clustering algorithm (DBSCAN [9]) on the check-in history of each user. The check-in points are then grouped into clusters each of which is in general of different size. We select the dominant cluster (say C_1), i.e., the one that includes the majority of the points, and we re-apply DBSCAN on C_1 to improve the estimation accuracy. Finally, we pick again the dominant cluster (say $C_{1,1}$) and we estimate the home location of the user as the centroid of the data points (lat/lon) in $C_{1,1}$.

2.2. Location 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**) [12]. When we refer to immutable characteristics it is clear that the main reason behind homophily is the mechanism of **selection** [11]. 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** [10] leads to homophily. With social influence, friendships are first created due to a number of plausible reasons - possibly unrelated with the characteristic under examination. Then people influence each other and become 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 [18]. 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 $\{(s, f): s \in S \land f \in F\}$ 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" [26].

If we further connect members of S based on their social relations, we obtain a *socio-affiliation* network (see Fig. 3). Using this structure 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. Based on the above, LBSNs that we consider in this work, can be modeled as a social-affiliation network where 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 to the specific spot. For instance, in Fig. 3, Bob has checked-into the "School of Information Sciences" and hence there is an affiliation edge that connects him with the corresponding focus. We refer to this network structure as *socio-spatial affiliation network*.

The socio-affiliation network as described above is a static network structure. In other words, it does not encode any information with regards to temporal information. Temporal information can further enable the study of fine-grained spatio-temporal behavior, by examining for instance the co-locations (in space and time) of users. While, in our dataset every check-in is timestamped, we choose to ignore this parameter in our analysis. The main reason for this is the fact the check-in time itself is not enough to accurately estimate the existence and duration of a co-location across time and space. This would require a *check-out* time as well. As we will discuss in more detail in Section 6 in more detail, there exist studies that utilize fine-grained temporal information (e.g., through cellular network data [25]). Our work is complementary to these studies. In particular, even if we do not know whether two friends' affiliations were created at the same time or with a time lag, their common affiliation is an indicator of a possible relation, and hence a socio-spatial tie. In this case the number of common affiliations might not be a strong predictor (as it has been shown to be the case when considering co-locations [5]), but as we will show there are other metrics computed on these common affiliations that are extremely informative.

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

2.3. Definitions

In this section we summarize the terminology/definitions used through the rest of the paper. In particular:

- **Location** or **Affiliation** or **Foci** refers to specific venues that appear in the Gowalla dataset.
- **Social edge** refers to a connection between two Gowalla users (i.e., represents declared, two-way, friendships between the users).

---

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

Table 1
Notations used.

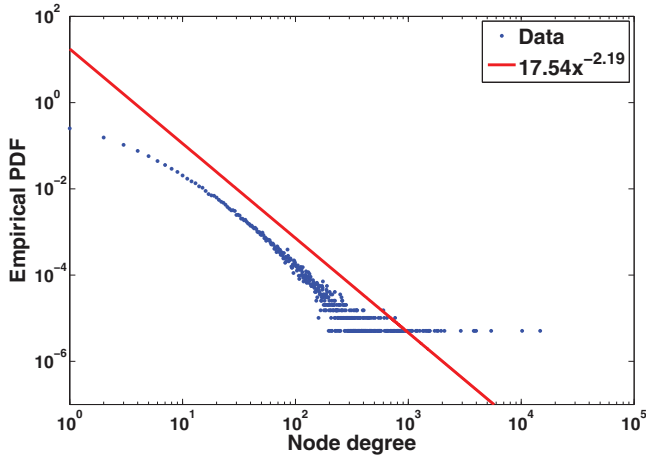|	
L_c	Set of venues that user c has visited/checked-in
$s(u, v)$	Similarity measure between users u and v
$P_l(u)$	Fraction of check-ins in affiliation l contributed by user u

Fig. 4. Node degree distribution.

- **Affiliation edge** or **User-Affiliation link** refers to a connection between a Gowalla user and an affiliation node (i.e., represents check-ins in the system).
- **Degree** or **Social degree** of a Gowalla user refers to the number of (social) edges that this user has.
- **Affiliation degree** of a location refers to the number of (affiliation) edges that the location has.

Finally, Table 1 provides a summary of the notations that will be used through the paper.

3. 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. In particular, we compute the empirical probability density function of a user's degree (Fig. 4). As we can observe, the degree distribution of Gowalla users exhibits a power law tail. This has been found to be true for other social networks as well [17], and implies that the majority of the users have very few friends, while very few users have many friends. Formally, the probability of a node u having a degree of k obeys the following rule:

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

We have also calculated the Maximum Likelihood Estimator for the power law exponent at the tail of the distribution. In particular, we have $\alpha = 2.1933$, while the statistical error of the estimation is $\sigma = 0.0057$. In Fig. 4 we have plotted the estimated degree distribution on top of the empirical PDF. Note that the fit is much better at the tail of the distribution as it is true for the vast majority of social networks. The average node degree is also computed to be 9.66.

Local clustering coefficient: Local clustering coefficient (cc for short) is tightly related to the notion of triadic closure. In particular, the (local) 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
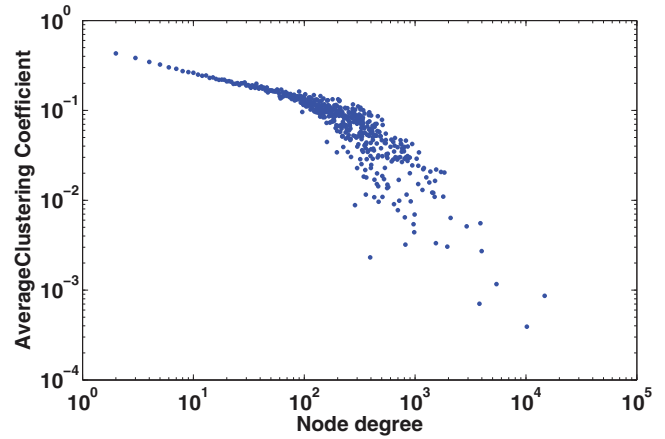

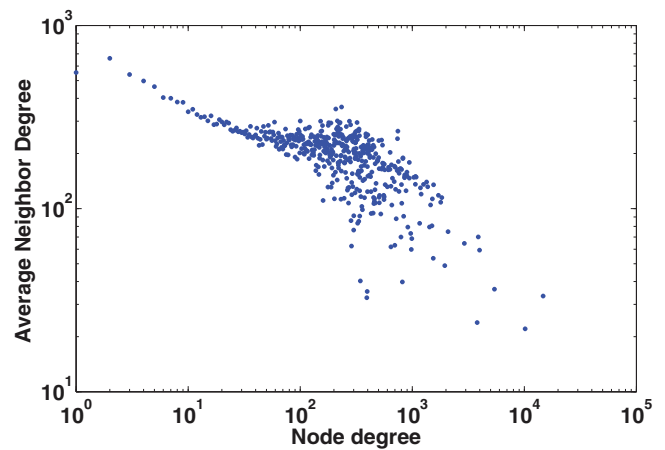**Fig. 5.** Clustering coefficient vs node degree.

Fig. 6. Average neighbor degree.

each other, over all the possible pairs between them, it is useful to be presented as a function of the node degree. Fig. 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 number is significantly higher from the expected 0.0491 clustering coefficient in a random network with the same degree distribution as Gowalla. The latter is computed using the random configuration model and is given by [17]

$$C_{rand} = \frac{1}{n} \frac{\left[\langle k^2 \rangle - \langle k \rangle\right]^2}{\langle k \rangle^3} \tag{2}$$

This transitivity, 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. Fig. 6 shows $d(k)$. 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 level of disassortativity; there are slightly more links connecting nodes of dissimilar degrees. This can be partially attributed to the actual structure of the network. In particular, there are a few only nodes with high degrees and hence there is a limited number of possible edges between high degree

Table 2
Quantiles of similarity for the different classes of user pairs.

-	80%	85%	90%	95%	99%
Friends	0.0192	0.0270	0.039	0.0636	0.16
Near non-friends	0	0.006	0.0115	0.02	0.04

nodes. Hence, the rest of the edges of these *hub*-vertices need to connect low and high degree nodes, leading to a slight disassortativity.

4. 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) [11]. In another study, Christakis and Fowler [3], a social network consisting of approximately 12,000 people, found that social relationships exhibit signs of homophily with regards to the obesity level. 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. In what follows, we take a first step towards examining homophily related to the locations visited by people. 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:

$$s(u, v) = \frac{|L_u \cap L_v|}{|L_u \cup L_v|} \quad (3)$$

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 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 latter pairs (of non-friends) as being in geographic proximity or not, based on their "home" locations. We have set up a threshold of 40 miles for defining pairs that are "nearby" or "distant".

Fig. 7 presents the cumulative distribution function (CDF) for $s(u, v)$ for pairs of friends and nearby pairs of non-friends. Clearly, friends have the highest similarity scores as compared to non-friends pairs even if the latter live within 40 miles from each other. Table 2 presents the quantiles of the corresponding empirical data. As we can see, 10% of the nearby friends have coefficients approximately 4%, which means that 4% of the places they have visited are common.
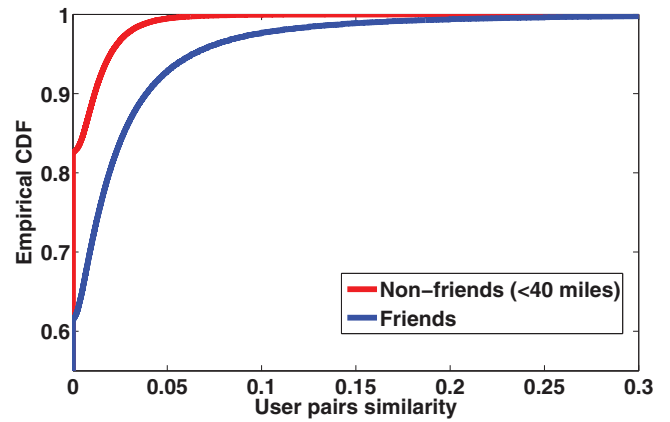

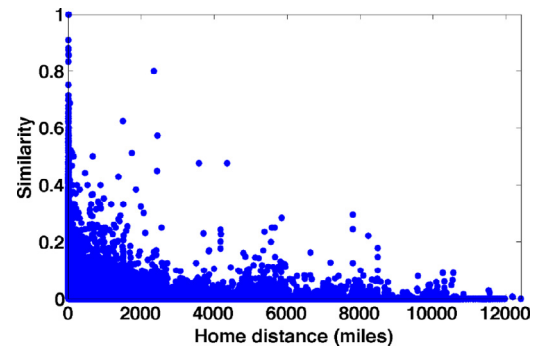**Fig. 7.** Empirical CDF for user similarity.

Fig. 8. Similarity tends to decrease with an increase in home distance.

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, in our friends pairs we have not distinguished them with regards to nearby or distant home locations. Even distant friends exhibit similarity 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, while users that are not friends, even if they are in proximity (e.g., in the same city) are unlikely to visit the same places.

Next, Fig. 8 shows the similarity values for friends as a function of the distance between home locations. Each point on the figure corresponds to a pair of friends. As we can see, there is a decreasing trend of the similarity value with distance between the friends' home location. This might have been expected, since people that live far away, even if friends, have less opportunities to "follow" the trails of each others. However, even for large distances (e.g., 1000 miles) the similarity is still non-negligible, since most possibly friends visit each other occasionally. Fig. 9 presents the same information for pairs of nearby non-friends. As we can see the similarity values are fairly stable over the span of the 40 miles considered, with a very slight decreasing trend. These values can be possibly attributed to common affiliations that users create because they live nearby (e.g., transportation hubs etc.) and not because of actual similarity in interests. It is interesting to note that friends living 2000 miles apart exhibit the same levels of similarity with non-friends that are within 40 miles.

Note here that in the definition of users similarity (Eq. 3), 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
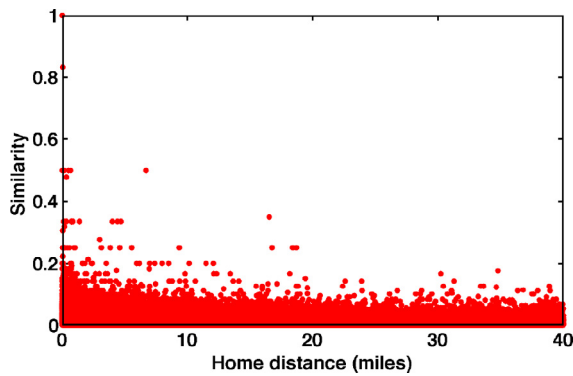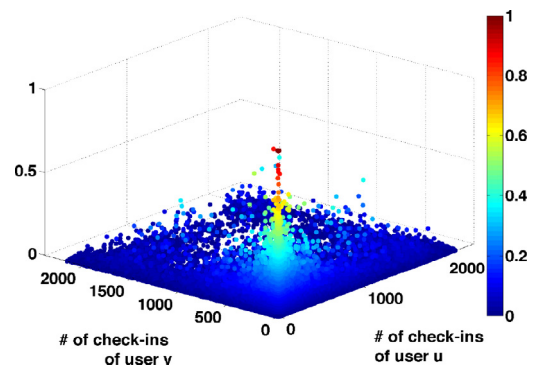
Fig. 9. Nearby non-friends' similarity.

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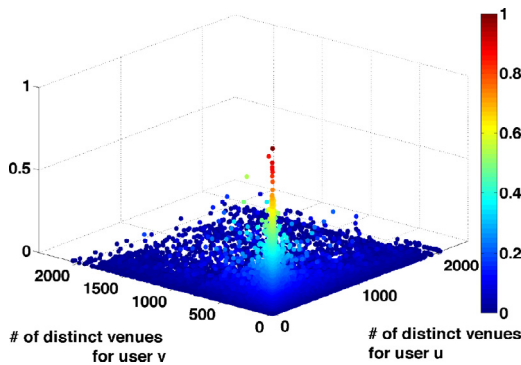

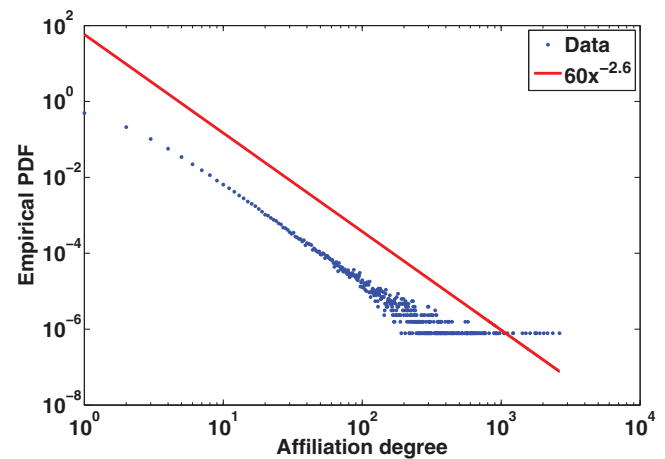**Fig. 10.** Similarity as a function of the users distinct affiliations.

Fig. 12. Affiliation degree distribution.

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 and discussed in Section 6. The importance of this finding is that it indicates that the characteristics of location information can be substantially different between friends and non-friends. We would also like to note here that there exist pairs of non-friends with high similarity values as well. For instance, there are 974,749 pairs ($\approx$ 1%) of non-friends that live within 40 miles, that have similarity values greater than 0.04! However, to reiterate, this high values for $s(u, v)$ can include "trivial" similarities, due to the fact that people *need* to visit some spots regardless of their actual preferences (e.g., airports, transportation hubs, etc.). In other words, all venues are treated equally in the definition of Eq. 3. Later, we will examine features/metrics of the spots visited from people that will be able to quantify how *accidental* a common affiliation can be considered.
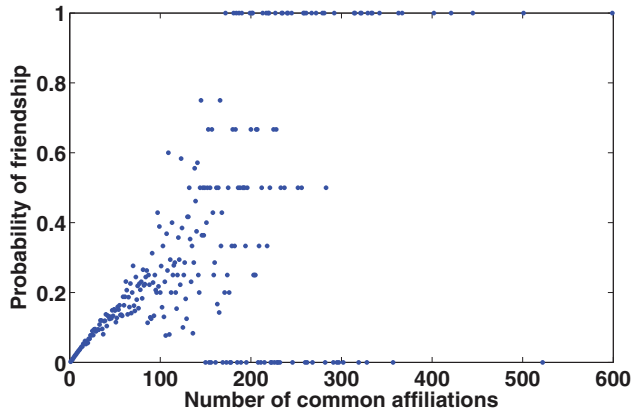
Figs. 10 and 11 present the similarity of two friends as a function of the number of their distinct affiliations and their check-in counts respectively. Even though our data consist of a static snapshot, this figure can be seen as an "emulation" of the temporal evolution of the similarity value of two 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 Eq. 3 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 5 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.

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. Fig. 12 shows the affiliation degree distribution, which as we can see exhibits a power law tail as well, with exponent $\alpha = 2.5953$ ($\sigma = 0.0067$). There are a few places with many visitors, while there are many venues with few visitors. The average focus degree is 3.11. Table 3 has the top and bottom 5 venues with regards to their degree. Note here that the bottom 5 venues were

Table 3
Top and bottom 5 venues based on their degree.

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

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.

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

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. Fig. 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
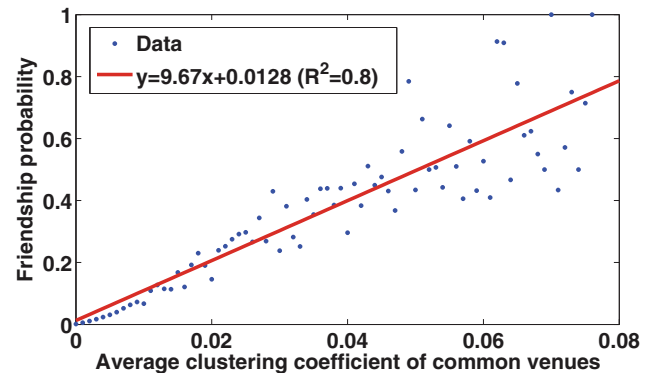
$$CC(l) = \frac{n}{k(k-1)/2} \qquad (4)$$

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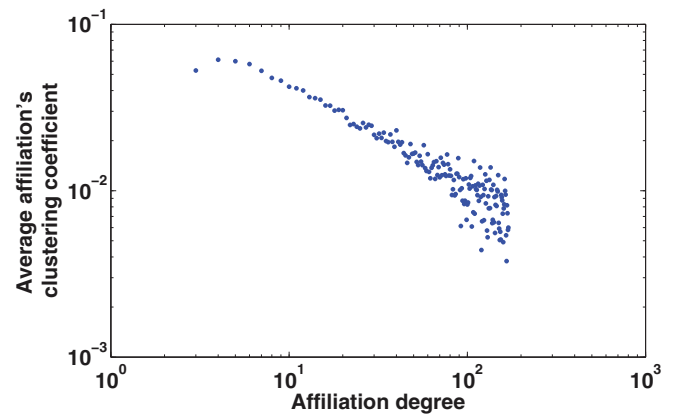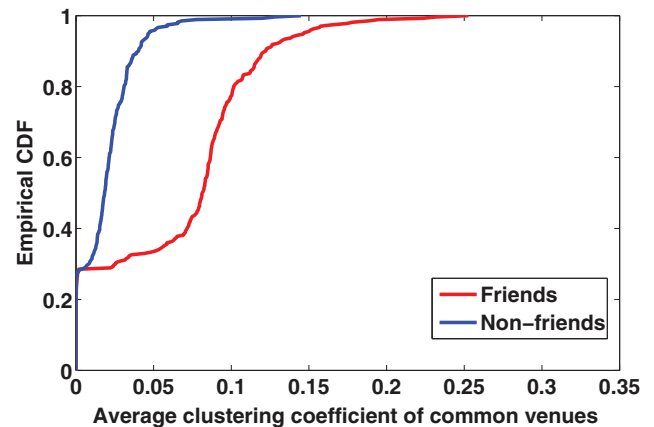
venue. The higher the cc is, the more connected are the people affiliated with it.

Fig. 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 Fig. 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, Fig. 16 plots the probability of friendship between two users as a function of the average
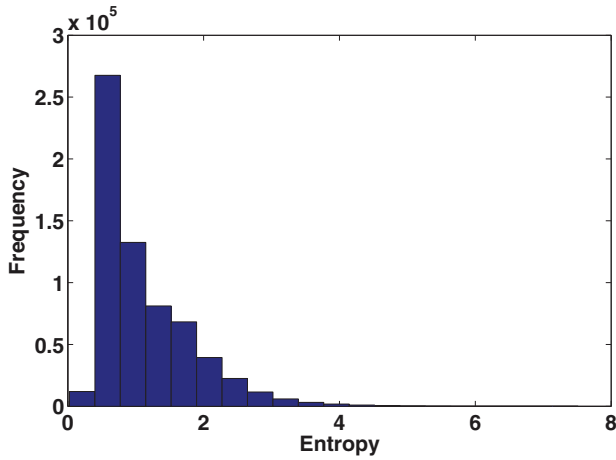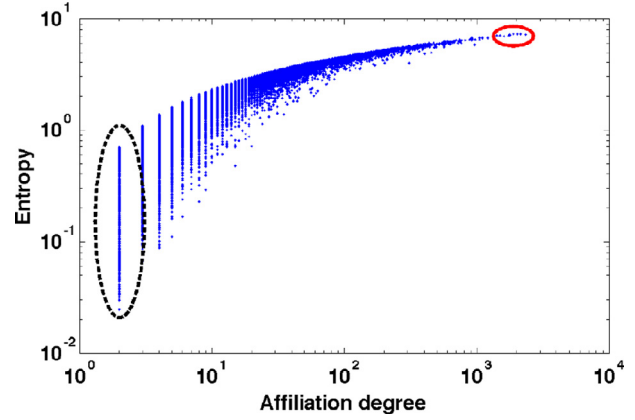
Fig. 17. The majority of the venus exhibit low entropy.

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

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

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 (calculated 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, large department store 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 (Fig. 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. [6] 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

$$e(l) = - \sum_{u:u \in S \wedge P_l(u) > 0} P_l(u) log(P_l(u)) \tag{5}$$

From Eq. 5 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 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.

Fig. 17 shows 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 (Fig. 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
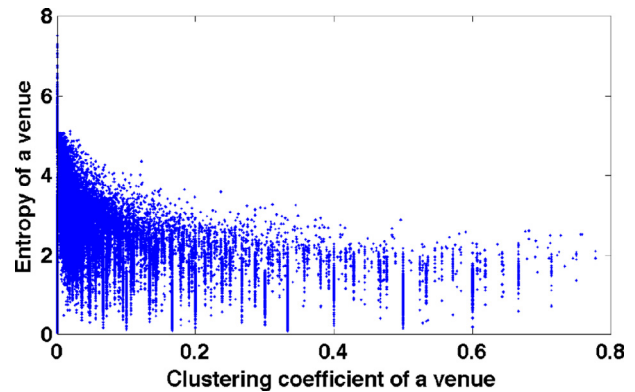
the dashed, black ellipse in Fig. 18, some venues with the lowest degree, do not exhibit the lowest entropy (although still smaller than 1).

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. Fig. 19 shows 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 Fig. 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 Fig. 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.
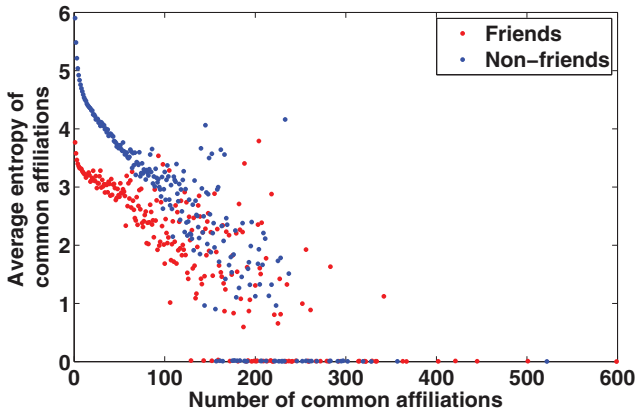
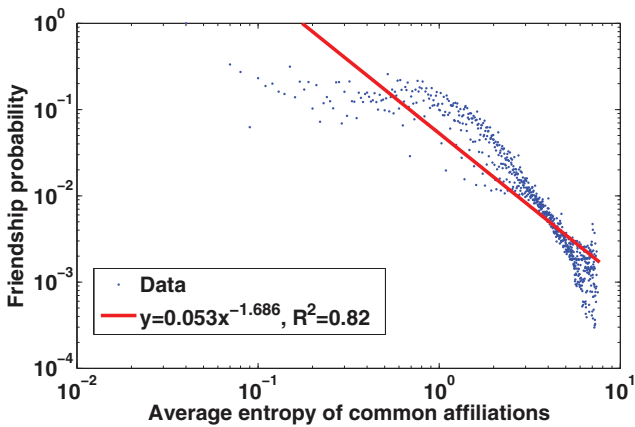**Fig. 20.** The common affiliations of friends exhibit lower (average) entropy.

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

Fig. 22. The predictive power of each considered feature using a simple, threshold-based, unsupervised learning algorithm.

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 Fig. 21 are promising. There appears to be a significant (negative) correlation between these two variables (correlation coefficient is equal to -0.7). Their relation is not linear as it was the case for the average cc of the common venues and the friendship probability (Fig. 16). Using least square we have fitted a power law with exponent $\alpha = -1.69$. 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 further examine this in the following section.

5. 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 4, 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
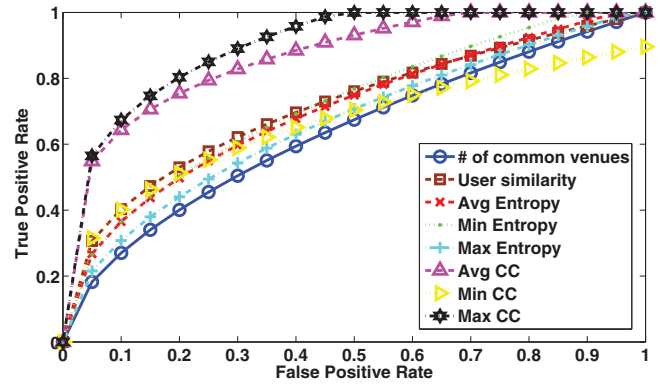
spatial/location information it might be possible to estimate future visits of users. The list of possible applications realized through the bonds between social and spatial information in an LBSN is long and not the focus of our study.

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 Fig. 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 classification algorithm on the possible affinities/ties of the social graph; we only want to examine the strength of the explored metrics in estimating social relations and showcase some aspects of the importance of our analysis.

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 Fig. 22. We focus on the positive (friends) instances, since there is a strong unbalanced distribution of the friends/non-friends instances in the network (only 3% of all possible social edges exist). Hence even a simple classifier that states every pair as non-friends, would exhibit a very good overall performance (i.e., 97% accuracy), 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 higher true positives on average). 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 two unsupervised clustering algorithm, that operates on the same set of features. First, we use a simple k-means algorithm for clustering. In particular, applying k-means clustering on each class (friends and non-friends) of our training set we compute r prototypes for each class ($r = 3$ in our case). The prototypes are essentially the centers of the clusters that k-means identified. Since in our case we have 2 classes, we have in total $2 \cdot r = 6$ labeled prototypes. In the test phase, we assign every test instance x to the class of the closest prototype. We evaluate the performance of this simple

Table 4
Performance of a simple *k*-means-based classifier.

$s(u, v)$	Min e	Max e	Avg. e	Min cc	Max cc	Avg cc	# foci	F-measure	Precision	Recall
✓	✓	✓	✓	✓	✓	✓	✓	0.36	0.48	0.29
✓	✗	✗	✗	✗	✗	✗	✗	0.35	0.51	0.27
✗	✓	✗	✗	✗	✗	✗	✗	0.56	0.52	0.6
✗	✗	✓	✗	✗	✗	✗	✗	0.52	0.56	0.49
✗	✗	✗	✓	✗	✗	✗	✗	0.54	0.51	0.57
✗	✗	✗	✗	✓	✗	✗	✗	0.32	0.55	0.22
✗	✗	✗	✗	✗	✓	✗	✗	0.47	0.69	0.35
✗	✗	✗	✗	✗	✗	✓	✗	0.45	0.70	0.33
✗	✗	✗	✗	✗	✗	✗	✓	0.33	0.49	0.25
✗	✓	✓	✓	✗	✗	✗	✗	0.53	0.51	0.57
✗	✗	✗	✗	✓	✓	✓	✗	0.44	0.76	0.31
✗	✗	✓	✗	✗	✓	✗	✗	0.51	0.54	0.49
✓	✗	✓	✗	✗	✗	✗	✗	0.52	0.56	0.49
✗	✗	✓	✗	✗	✗	✗	✓	0.31	0.34	0.28

Table 5
Performance of an svm classifier with linear kernel.

$s(u, v)$	Min e	Max e	Avg. e	Min cc	Max cc	Avg cc	# foci	F-measure	Precision	Recall
✓	✓	✓	✓	✓	✓	✓	✓	0.68	0.90	0.54
✓	✗	✗	✗	✗	✗	✗	✗	0.47	0.75	0.34
✗	✓	✗	✗	✗	✗	✗	✗	0.67	0.70	0.65
✗	✗	✓	✗	✗	✗	✗	✗	0.56	0.66	0.49
✗	✗	✗	✓	✗	✗	✗	✗	0.65	0.72	0.59
✗	✗	✗	✗	✓	✗	✗	✗	0.39	0.84	0.26
✗	✗	✗	✗	✗	✓	✗	✗	0.65	0.90	0.50
✗	✗	✗	✗	✗	✗	✓	✗	0.65	0.88	0.51
✗	✗	✗	✗	✗	✗	✗	✓	0.23	0.69	0.14
✗	✓	✓	✓	✗	✗	✗	✗	0.67	0.71	0.64
✗	✗	✗	✗	✓	✓	✓	✗	0.66	0.89	0.52
✗	✗	✓	✗	✗	✓	✗	✗	0.65	0.90	0.51
✓	✗	✓	✗	✗	✗	✗	✗	0.63	0.78	0.53
✗	✗	✓	✗	✗	✗	✗	✓	0.62	0.71	0.55

supervised classifier through precision, recall and *f*-measure:

$$precision = \frac{TP}{TP + FP} \tag{6}$$

$$recall = \frac{TP}{TP + FN} \tag{7}$$

$$f - measure = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{8}$$

where TP stands for true positive, FP stands for false positive and FN stands for false negative. 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 *f*-measure is the harmonic mean of precision and recall and hence, takes high values when both are high. If one of the two is low, the *f*-measure will tend to be low as well. Therefore, the *f*-measure can be thought as evaluating the overall performance of the classifier. Using 10-fold cross validation we obtain the results presented in Table 4. Each row corresponds to a different set of features used for the classification. As we can see again, when we include in the features attributes related with the entropy or the cc of the common affiliations the corresponding *f*-measure is higher. For example, by using the average entropy of the common venues visited by two people we are able to recover 49% of the actual friendships, while from all our friendship predictions, 56% of them are correct (*f*-measure=0.54). When using only the number of common affiliations we obtain a value for the *f*-measure of only 0.33. We have further examined an svm classifier with linear kernel on exactly the same combinations of features (see Table 5). As one might have expected the performance is further improved compared to the simple *k*-means classifier.

To summarize, 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 social relationship inference, location prediction etc. While our study does not aim into providing a state-of-the-art social link classifier, we showed that even with simple schemes one can achieve very good performance. More advanced algorithms can further boost the performance, but it is beyond the scope of our work.

6. Related work

The current work extends our previous study [22]. Compared to [22] we have improved the estimation of the home location of the users in our dataset in finer grain granularity and we have updated the relevant results. We have also improved the social network analysis (utilizing MLE for estimating the power exponent and calculating degree assortativity). Finally, we have delved into the details of the supervised classification algorithm that utilizes *k*-means, while we have also added new performance results of an SVM classifier with a linear kernel.

Furthermore, 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. [20] 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 [19]. In alignment, Cho et al. [2] use cell phone location and LBSN data to understand the laws dictating human mobility. Li and Chen [13] 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. [23] try to use information from both components to identify the relation between friendship and geographic distance using data from three different LBSNs (Gowalla, Foursquare and Brightkite). They find that the socio-spatial structure of these systems cannot be explained only by geographic factors or only by 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. [21] obtain a static snapshot of Foursquare in order to analyze the activity in different neighborhoods of London and New York, while Ye et al. [27] 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. [7] [8], as well as Li et al. [14], 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 [16], simply accounting for the number of co-existences can be expected to not be very accurate. Recently, Wang et al. [25] using mobile phone data have identified a positive correlation between mobile homophily, network proximity and social tie strength. Furthermore, Crandall et al. [5], using data from Flickr developed a simple model for quantifying the probability of friendship between two users given the number of their co-locations. Cranshaw et al. [6] 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. Scellato et al. [24] 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. Finally, Zhang et al. [28] study the mechanisms of homophily and social selection in the context of socio-spatial affiliation networks.

To reiterate, our work is complementary to all the aforementioned studies. First, as it might be evident from the discussion above, the related studies deal with slightly different problems (e.g., friend and/or location recommendations, analysis of structural properties of LBSNs etc.). More importantly though, the studies that are closer to our work (i.e., [5,25]) consider also fine-grained temporal information. For instance, the authors in [5] show that the number of co-locations are extremely informative of the social relation of people, while Cranshaw et al. [6] compute features such as the duration of a co-location. Wang et al. [25] have a very detailed mobility dataset as well (Call Detail Records) from which they can accurately and with high confidence extract the co-locations in time. On the contrary, in our study, we consider only spatial information, through affiliations of users with specific venues. We do not regard the temporal dimension, since in many cases co-occurrence in time might be hard to establish with certainty as explained in Section 2. Then we delve into the details of these common affiliations/locations. We examine many different (network) metrics (e.g., number, average entropy, average clustering coefficient etc. of common affiliations) and the additional information each one of these can provide with regards to the underlying social structure. We further believe that the fact that some of the results reported in these studies appear to also hold with our different dataset supports the conclusion that there are some strong underlying bonds between mobility and social connections, hence, complementing these results.

7. Discussion and limitations

While our analysis and models have been based on a dataset obtained from Gowalla, we believe that the intuitive interpretation of our results make them generic enough. Nevertheless, we acknowledge that there are specific steps in our analysis that can lead to inaccuracies. For example, given the absence of ground truth for the home location of a user we iteratively apply DBSCAN on the check-ins of each user. The iterative process aims at reducing the estimation error. However, the method will fail if users choose to check-in mainly during trips (e.g., as a means of keeping a travel diary). Apart from such cases, which we expect to be a small portion of the user base, our analysis is based on a coarse-grain view of the distance between two users - e.g., at the city level. Therefore, even if there will definitely be errors during this estimation process, we do not expect them to significantly alter our conclusions/models.

Furthermore, during the computation of the users similarity $s(u, v)$ we did not consider the temporal dynamics. In particular, one can argue that if both users u and v visited location l but with a large time-gap, this should not contribute equally to their estimated similarity as compared to check-ins that happened with a small time-gap (e.g., within the same day). While, this is in principle true, applying it to our dataset might be problematic and this is the major reason we chose not to consider the temporal dimension. In particular, given the voluntary nature of check-ins, users very often do not check-in to venues every time they visit the location [4,15]. This can potentially lead to a significant underestimation of the similarity when incorporating time in the computation of $s(u, v)$ and consequently also affect our models and conclusions.

8. Conclusions

In this paper we model an LBSN as a socio-spatial 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 percentage of common venues visited, as compared to non-friends. However, 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 components better. This is also supported by the evaluations and results from simple, social link classifiers.

References

[1] Z. Cheng, J. Caverlee, K. Lee, D. Sui, Exploring millions of footprints in location sharing services, in: ICWSM, 2011.

[2] E. Cho, S.A. Myers, J. Leskovec, Friendship and mobility: Friendship and mobility: user movement in location-based social networks, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (ACM KDD), 2011, pp. 279–311.

[3] N. Christakis, J. Fowler, The spread of obesity in a large social network over 32 years, N. Engl. J. Med. 357 (4) (2007) 370–379.

[4] H. Cramer, M. Rost, L. Holmquist, Performing a check-in: emerging practices, norms and conflicts in location-sharing using foursquare, in: ACM MobileHCI, 2011.

[5] D. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, J. Kleinberg, Inferring social ties from geographic coincidences, in: Proceedings of the National Academy of Sciences, 2010, pp. 159–178.

[6] J. Cranshaw, E. Toch, J. Hong, A. Kittur, N. Sadeh, Bridging the gap between physical location and online social networks, in: Proceedings of the 12th ACM international conference on Ubiquitous computing (UBICOMP), 2010.

[7] N. Eagle, A. Pentland, Eigenbehaviors: identifying structure in routine, in: Behavioral Ecology and Sociobiology, 2009.

[8] N. Eagle, A. Pentland, D. Lazer, Inferring friendship network structure by using mobile phone data, in: Proceedings of the National Academy of Sciences, 2009.

[9] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (ACM KDD), 1996.

[10] N. Friedkin, A Structural Theory of Social Influence, in: Cambridge University Press, 1998.

[11] D. Kamdel, Homophily, selection and socialization in adolescent friendships, Am. J. Sociol. 84 (2) (1978) 427–436.

[12] P. Lazarsfeld, R. Merton, Friendship as a social process: a substantive and methodological analysis, in: Freedom and Control in Modern Society, Van Nostrand, 1954.

[13] N. Li, G. Chen, Analysis of a location-based social network, in: IEEE CSE, 2009.

[14] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, W. Ma, Mining user similarity based on location history, in: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS), 2008.

[15] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, J. Zimmerman, Im the mayor of my house: Examining why people use foursquare - a social-driven location sharing application, in: ACM CHI, 2011.

[16] A. Miklas, K. Gollu, K. Chan, S. Saroiu, K. Gummadi, E.D. Lara, Exploiting social interactions in mobile systems, in: UbiComp, 2007.

[17] M. Newman, Networks: An Introduction, Oxford University Press, 2010.

[18] M. Newman, D. Watts, S. Strogatz, Random graph models of social networks, in: Proc. Natl. Acad. Sci. USA, 2002.

[19] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, C. Mascolo, A tale of many cities: universal patters in human urban mobility, PLoS ONE 7 (5) (2011a) e37027, doi:10.1371/journal.pone.0037027.

[20] A. Noulas, S. Scellato, C. Mascolo, M. Pontil, An empirical study of geographic user activity patterns in foursquare, in: ICWSM (poster session), 2011b.

[21] A. Noulas, S. Scellato, C. Mascolo, M. Pontil, Exploiting semantic annotations for clustering geographic areas and users in location-based social networks, in: SMW, 2011c.

[22] K. Pelechrinis, P. Krishnamurthy, Trust representation and aggregation in a distributed agent system, in: Proceedings of the ECML/PKDD, 2012.

[23] S. Scellato, A. Noulas, R. Lambiotte, C. Mascolo, Socio-spatial properties of online location-based social networks, in: Proceedings of ICWSM, 2011a.

[24] S. Scellato, A. Noulas, C. Mascolo, Exploiting place features in link prediction on location-based social networks, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (ACM KDD), 2011b.

[25] D. Wang, D. Pedreschi, C. Song, F. Giannotti, A.-L. Barabasi, Human mobility, social ties, and link prediction, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining(ACM KDD), 2011.

[26] D. Watts, Six Degrees: The Science of a Connected Age, W.W. Norton & Company, 2003.

[27] M. Ye, P. Yin, W.-C. Lee, Location recommendation for location-based social networks, in: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS), 2010.

[28] K. Zhang, K. Pelechrinis, Understanding spatial homophily: the case of peer influence and social selection, in: Proceedings of the 23rd International Conference on World wide web (ACM WWW), 2014.