

Understanding Spatial Homophily: The Case of Peer Influence and Social Selection

Ke Zhang
School of Information Sciences
University of Pittsburgh
kez11@pitt.edu

Konstantinos Pelechris
School of Information Sciences
University of Pittsburgh
kpele@pitt.edu

ABSTRACT

Homophily is a phenomenon observed very frequently in social networks and is related with the inclination of people to be involved with others that exhibit similar characteristics. The roots of homophily can be subtle and are mainly traced back to two mechanisms: (i) social selection and (ii) peer influence. Decomposing the effects of each of these mechanisms requires analysis of longitudinal data. This has been a burden to similar studies in traditional social sciences due to the hardness of collecting such information. However, the proliferation of online social media has enabled the collection of massive amounts of information related with human activities. In this work, we are interested in examining the forces of the above mechanisms in the context of the locations visited by people. For our study, we use a longitudinal dataset collected from Gowalla, a location-based social network (LBSN). LBSNs, unlike other online social media, bond users' online interactions with their activities in real-world, physical locations. Prior work on LBSNs has focused on the influence of geographical constraints on the formation of social ties. On the contrary, in this paper, we perform a microscopic study of the peer influence and social selection mechanisms in LBSNs. Our analysis indicates that while the similarity of friends' spatial trails at a geographically global scale cannot be attributed to peer influence, the latter can explain up to 40% of the geographically localized similarity between friends. Moreover, this percentage depends on the type of locations we examine, and it can be even higher for specific categories (e.g., nightlife spots). Finally, we find that the social selection mechanism, is only triggered by places that exhibit specific network characteristics. We believe that our work can have significant implications on obtaining a deeper understanding of the way that people create friendships, act and move in real space, which can further facilitate and enhance applications such as recommender systems, trip planning and marketing.

Categories and Subject Descriptors

J.4 [Social and Behavioral Sciences]: Sociology; H.2.8 [Database Applications]: Data mining

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
WWW'14, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2744-2/14/04.
<http://dx.doi.org/10.1145/2566486.2567990>.

General Terms

Social Network Analysis, Data Analytics, Data Mining Methods

Keywords

Homophily; Peer Influence; Social Selection; Location-based Social Networks

1. INTRODUCTION

Homophily - also referred to as assortative mixing - is a phenomenon that appears very often in (social) networks. A (positively) mixed network, is one where the number of ties/edges between vertices that exhibit the same characteristics is significantly higher compared to the number that would have been expected if connections were made at random. McPherson *et al.* [14] refer to this phenomenon as “the birds of a feather flock together”, and present many instances of homophily in social networks with regards to a large spectrum of people's attributes (e.g., age, religion, education, occupation, behavior etc.).

While mixing patterns in a network with respect to a specific characteristic can be formally and precisely quantified (e.g., assortativity coefficient [16]), the reasons behind their existence are not clearly understood and might differ for different scenarios. Nevertheless, there are two general mechanisms that are usually cited as the roots of homophily: (i) peer influence and (ii) social selection.

Peer Influence: Peer influence appears specifically when we examine mutable characteristics, such as behavior, political views etc. When this mechanism is in play, people first become friends for reasons that are possibly not related to the characteristic X under examination, and then one *influences* the other on decisions related to X . In this paper, we are interested in studying mixing patterns with regards to locations visited by people (as we will see in detail in Section 3 these patterns are homophilous). Given that this is a mutable characteristic, peer influence can be a possible cause of the observed assortativity. Figure 1 illustrates the peer influence mechanism in the context examined in our work. In this figure, we depict a socio-affiliation network, where affiliations (shown as the rectangular nodes) are the actual venues that people visit. We have further timestamped representative edges, with the time of their creation. In particular, Joe and Jack became friends at time t_k , while Joe has visited “Li's Restaurant” at time t_{k-n} (that is, prior to becoming friends with Jack). On the other hand, Jack has not visited “Li's Restaurant” prior to time t_k . Assuming peer influence between Joe and Jack, an affiliation edge between Jack and “Li's Restaurant” will appear some time after they become friends (e.g., t_{k+m}) as presented in the figure. Simply put, when peer influence operates, people tend to first form a social tie and then become (more) similar.

Social Selection: Social selection can cause assortative mixing in networks, either with regards to mutable or immutable (e.g., age, race, sex etc.) characteristics. When social selection acts, people tend to associate with others that are already similar to them with regards to the characteristic under examination. In other words, people are already similar and this is essentially the cause of the friendship creation. Figure 2 illustrates the above concept. As we see, Joe and Alice, became friends at time t_l . Prior to that, they exhibit a large similarity with regards to the places visited, since they both visited “Li’s Restaurant” and “Mike’s Coffee Shop”. Simply put, when social selection operates people first become (or are by nature) similar and then they create a social tie.

To reiterate, both of the above processes lead to the same observable phenomenon, that is, homophily with regards to the locations visited by friends in our setting. However, it might be hard to trace back to its actual roots. As it should be obvious from the above, one needs longitudinal data in order to decompose the reasons behind assortative mixing. This has traditionally been a burden for large scale studies on this topic. However, during the last years there is a rapid penetration of online social media in people’s daily activities. This, in turn, has enabled the collection of massive datasets that can foster social studies on human interactions. For instance, Bakshy *et al.* [3] using data collected from Twitter, examine the way that people adopt and share content, while Goel *et al.* [10] study how content diffuses through the underlying social network.

In this paper, we use a longitudinal dataset obtained from Gowalla, a location-based social network, in order to examine the reasons behind the homophilous patterns observed with regards to the actual spots visited by people. In particular, we study the existence as well as the levels of peer influence and social selection in the network. Our approach is microscopic, in the sense that we consider the above mechanisms in a variety of granularities. In particular, we consider global versus local influence, the relation between the actual type of location and the underlying peer influence, as well as the impact of the type of location on the *effectiveness* of the social selection process. Our main findings can be summarized in the following:

- While the similarity of users’ geo-trails at a global scale cannot be attributed to peer influence, the latter can explain on average up to 40% of the geographically local similarity between friends.
- The levels of local peer influence differ depending on the type of the location we consider.
- The social selection mechanism works upon non-trivial similarity and can be stimulated by specific types of venues.

Disclaimer: We would like to emphasize from the beginning of our study, that the conclusions drawn from our analysis are essentially limited by the extend to which the Gowalla dataset represents the real-world activities of its users. Hence, while for ease of presentation we might use general phraseology for our conclusions, we acknowledge that the latter might exhibit biases. Nevertheless, this is an essential limitation of every computation sociology study, which is based on digital trails of people obtained from online social media/platforms.

Roadmap: The rest of the paper is organized as follows. Section 2 discusses related literature and describes the longitudinal dataset we use for our analysis. In Section 3 we examine the existence of homophily with regards to the locations visited by people. Section 4 provides a detailed, microscopic analysis of peer influence, while Section 5 examines the social selection mechanism.

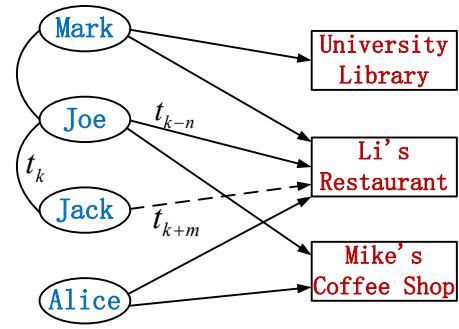


Figure 1: Peer influence mechanism.

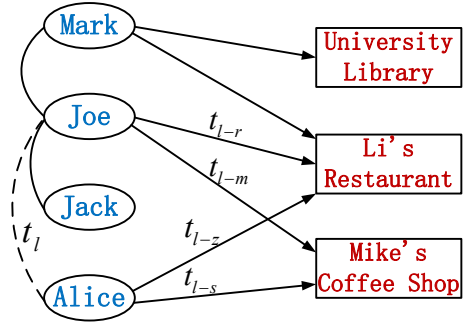


Figure 2: Social selection mechanism.

Finally, Section 6 briefly discusses the scope and limitations of our study and concludes our work.

2. RELATED STUDIES AND EXPERIMENTAL SET UP

In this section we begin by briefly discussing studies related to our work. Later we will introduce our dataset and establish our experimental set up.

2.1 Related Studies

As aforementioned the availability of electronic traces of human activities has enabled the study of topics related to homophily, information diffusion, social selection and peer influence. For instance, Kossinets and Watts [12] utilize a dataset comprising of e-mail communications between members of a large U.S. university to study the origins of the homophily in the underlying communication network. The same authors [11] study the effects of social selection on friendship formation using again an e-mail communication network of a U.S. university. They show that the friendship probability between two students increases up to a certain number of common interests - as captured from the number of common classes between students - and then remains constant. In the same direction, Lewis *et al.* [13] examine the mechanisms of social selection and peer influence in a group of students of a U.S. institution by studying the co-evolution of the friendships and tastes in music, movies and books, over a period of four years.

Other studies provide some basic intuition on how innovations propagate through the network and they further show that social ties can facilitate information contagion and consequently influence users’ actions. For instance, Bakshy *et al.* [4] use data from Sec-

and Life to examine the social influence on the adoption of content by the users. Among other findings, they show that adoption rate increases with an increased number of friends that have already adopted a content. However, as one might expect not all adoptions can be attributed to peer influence and in-network effects. Myers *et al.* [15] studied the effect of external influences on information diffusion using data collected from Twitter. They further developed a model through which they can quantify external influence over time. Very recently, Tang *et al.* [21] studied conformity, which can be thought of as a special type of social influence. In particular, conformity is the action of matching one's actions to the norms of the groups he belongs to. The authors' results indicate that conformity exists in all four digital social network datasets they examined.

A different line of work, studies statistical methodologies for unveiling the roots of homophily. For instance, Anagnostopoulos *et al.* [1] design a statistical test, the shuffle test, for deciding whether peer influence is a likely source of the observed homophily. In brief, the key idea behind the shuffle test, is that if influence is not a possible source of the assortative mixing, timing of actions should not matter. Hence, *reshuffling* of the timestamps of the events, should not significantly change the assortativity level in the network. Another statistical framework for distinguishing between influence and selection effects in dynamic networks was presented in [2]. In particular, the authors develop a dynamic matched sample estimation framework and apply it on a large-scale network dataset that captures the adoption of a specific product. In parts of our work, we use an approach similar to that followed by Crandall *et al.* [6]. The authors study the social selection and influence in online communities such as Wikipedia and LiveJournal. In order to distinguish between social selection and influence, they examine the temporal evolution of the friends similarity prior and after they became associated.

In our work, we focus on a novel type of online social media, namely LBSNs, that only recently has attracted attention from the research community. The key difference between traditional online social media and LBSNs, is that the latter directly relates online interactions with physical space, and hence, studies of LBSNs can have stronger implications of actual real-world behaviors of people. The work from Cho *et al.* [5] is the closest one to our study. In particular, the authors examine the relationship between friendship and users' mobility. They show that the actual geographic location (latitude/longitude) that users travel to is influenced by the presence of a friend or not. They further show that this influence increases with an increase in the distance between the home locations of friends. In our work we further delve into the details of peer influence, and in particular we examine both its geographic scope, as well as its contextual properties (existing literature seems to support the general connection between influence and *topics* [20]). In particular, we do not simply consider geographic locations, but actual venues, and examine the strength of influence for different types of places. Furthermore, we examine the social selection mechanism and how this is realized through the different types of venues (e.g., are there specific characteristics of the venues that *promote* friendship creation by stimulating social selection?).

2.2 Our Dataset and Analysis Set Up

The longitudinal dataset that we used for our study was provided to us by the authors of [19]. It was crawled from Gowalla, a commercial LBSN¹, between May 05, 2010 and August 18, 2010. The dataset consists of 10,097,713 public check-ins performed by 183,709 users in 1,470,727 distinct places. Every venue is associ-

ated with a category, that essentially describes the type of location the user checked-in. There are 283 distinct categories in Gowalla. Every check-in log is a tuple of the form $\langle \text{User ID}, \text{Venue ID}, \text{Latitude}, \text{Longitude}, \text{Time}, \text{Category ID} \rangle$. 27,895 venues are unclassified (i.e., $\text{Category ID} = \text{NULL}$) and hence, we discard them. This results in a dataset with 10,062,916 check-ins, in 1,442,832 distinct places by 183,500 users (209 users had check-ins only in unclassified spots).

Gowalla users also participate in a friendship network with reciprocal relations, which consists of 765,871 links. Gowalla was crawled every day for the aforementioned period, and hence the formation time of every friendship that was created after May 05, 2010 was able to be obtained. For the purposes of our work, we will use only the pairs of friends for which we have the actual friendship creation time. There are 289,888 such links in total. Some of the edges may also have been deleted (e.g., Jack “de-friends” Alice). For these links we also have a deletion time. Hence, the friendship edges have the following 4-tuple form $\langle \text{User ID}, \text{Friend ID}, \text{Formation Time}, \text{Deletion Time} \rangle$. From the 289,888 links above, only $\approx 2\%$ of them were deleted afterwards, and thus, we can safely discard them. If we further keep only pairs of friends for which both users have at least one check-in we have a final number of 202,424 links that we use for our study.

Home Location of a User: Our dataset does not include home location information for the users. However, we are interested in examining the mechanisms of peer influence and social selection in relation to the distance between the home locations of the users. In order to infer the home locations of the users, 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 with the maximum number of the data points (i.e., check-ins), 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 $C_{1,1}$.

Definitions: Before moving on to our analysis, we wish to introduce some terminology that we will be used throughout the rest of the paper. Two users u and v are said to have been **check-in co-located**, if they have both checked-in to at least one common venue, regardless of the actual check-in time. Furthermore, u and v are said to have been **area co-located at v 's home location**, if u has checked-in to at least one venue, within 25kms from v 's home location.

In the above definitions, we do not impose a constraint of co-location both in time and space². When Alice and Bob influence each other with respect to some behavior, e.g., adopting a specific product, it does not necessarily mean that they will buy it at the same time. This is exactly what Figure 1 depicts. Furthermore, similarity is related with the actual actions and not necessarily when these actions take place (e.g., two people that write on a specific Web blog can still be considered similar, regardless of whether they both write on Thursday nights or not.).

3. SPATIAL HOMOPHILY

Traditionally, vertices in a network are annotated with scalar or enumerative characteristics and metrics for quantifying the level of homophily in these scenarios are very well defined [16]. Nevertheless, in our case, we want to evaluate the mixing patterns in the network with regards to the spatial behavior of users, that is, the

¹Gowalla has been acquired from Facebook and ceased its operations in March 2012.

²Note that this would require knowledge of a “check-out” time as well.

places they visit, which cannot be described by a single number or label.

A user u of our LBSN is associated with a vector \mathbf{c}_u capturing the places he has visited. In particular the i^{th} element of vector \mathbf{c}_u , is equal to the number of check-ins that u has in venue i . Since we cannot directly compare vectors and directly apply the assortativity coefficient [16], we rely on a different methodology. For our purposes, we will need to define a similarity measure between vectors. In this work, we will utilize the cosine similarity. In particular, the similarity between two users u and v is defined as:

$$sim_{u,v} = \frac{\mathbf{c}_u \cdot \mathbf{c}_v}{\|\mathbf{c}_u\|_2 \|\mathbf{c}_v\|_2} \quad (1)$$

In order to identify the existence of assortative mixing - or not - in the network we will follow the same line of thought as in the definition of the assortativity coefficient, tailored though in our context. The assortativity coefficient essentially estimates the difference between the actual number of edges in the network that fall between vertices of the same type (enumerative characteristic) or of similar attribute value (scalar characteristic) and those that would have been expected if connections were made at random. Adopting this idea in our context, we ought to calculate the average spatial similarity between friends in the real network, sim_{real} , and compare it with the expected average similarity if connections were made at random. In order to calculate the latter we will rely on Monte Carlo simulations. In particular, we will sample the ensemble of $G(n, m)$ Erdős-Rényi random graphs³ and calculate the average spatial similarity between friends in the sampled networks, sim_{rnd} . If $sim_{rnd} \ll sim_{real}$, the network essentially exhibits homophily.

Nevertheless, sampling the pure $G(n, m)$ model might lead to under-estimation of the average similarity value. In particular, it has been found that the majority of one's friends live in nearby locations [18] [5]. In other words, the probability distribution of the home location distance between two friends d_f has the majority of its mass concentrated into small distances. We have also verified this is true in the dataset we are using (see Figure 3(a)). This can have implications on the sim_{rnd} value as computed above. In particular, since the majority of the user pairs live far from each other (the number of pairs of users living in the same city is much less compared to all possible pairs of users), $G(n, m)$ sampling will lead to edges between users that live far away. Such pairs though are also expected to have much lower similarity, since they simply do not have many chances to visit the same places. Hence, we also perform a second series of Monte Carlo simulations, where we sample from a modified, location-aware, $G(n, m)$ ensemble. In particular, we pick the first end of an edge uniformly at random, and we use the distribution of d_f to randomly select the other end of the edge. In other words, while we randomly sample the edges, we make sure to preserve the distribution of the friends' home distance. Using these randomized networks we can calculate the average similarity between friends, $sim_{l,rnd}$.

We sample the two randomized networks 100 times and then calculate the 95% confidence interval (CI) for the average similarity between friends. Our results are presented in Table 1. As we can see the value of the friends' average similarity in the real network lays outside the 95% confidence intervals for both random network models and is significantly higher as compared even to the upper bound of these CIs. This leads us to the conclusion that the network under consideration exhibits strong assortative patterns with

³A brief background on Erdős-Rényi random graphs is provided at Appendix A.

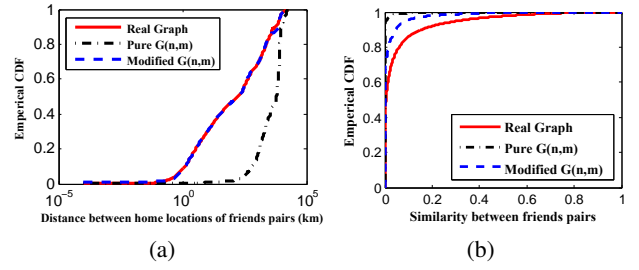


Figure 3: (a) Our modified random graph ensemble retains the distribution of home location distances observed in the real network. (b) Similarity between friends in a real network is much higher compared to that in the randomized networks.

respect to the spatial trails of the users. This strengthens the results reported by Wang *et al.* [22] where a correlation between spatial trajectory similarity and network closeness is reported using call detail records as well as those in [17], where a different, less rigorous, method was used along with a different similarity metric between users. Finally, Figure 3(a), presents the cumulative distribution function of the home distance between two friends for the real network, a pure $G(n, m)$ representative sample and a modified $G(n, m)$ representative sample. As we can see the spatially modified $G(n, m)$ model exhibits a similar home distance distribution with the one of the real network. On the contrary the pure Erdős-Rényi random graph exhibits longer home distances overall as expected. Figure 3(b) further depicts the cumulative distribution function of the similarity values for the connected vertices in the real network, and the representative random graph samples used in Figure 3(a). Results verify the ones we obtained in Table 1.

Table 1: There is a clear homophily with regards to the spatial trails of Gowalla users.

sim_{real}	$sim_{l,rnd}$	sim_{rnd}
0.05425	[0.01836, 0.01837]	[0.00236, 0.00237]

4. PEER INFLUENCE

Having established the existence of spatial homophily in the network we turn our attention to decomposing the reasons behind this phenomenon. In this section, we examine the peer influence mechanism with regards to spots visited by people. Our analysis considers both (i) the geographical scope of peer influence (i.e., whether people are influenced at a global/local scale) and (ii) the context of peer influence (i.e., whether people are influenced - or not - at the same degree with regards to different types of places).

4.1 Global Influence

We begin by examining the **global influence** between people. By the term global, we essentially refer to possible effects people can have on their friends' decisions related with their check-ins at *any* part of the world. In other words, if Bob, who is from New York City, and his friend Alice, who is from Boston, visited "Restaurant X" in San Francisco, was it a result of peer influence between each other? We would like to emphasize here that, while from a sociological perspective the question of whether peer influence affects the check-ins of a pair of friends anywhere in the world might seem absurd, we begin with this question in order to smoothly introduce

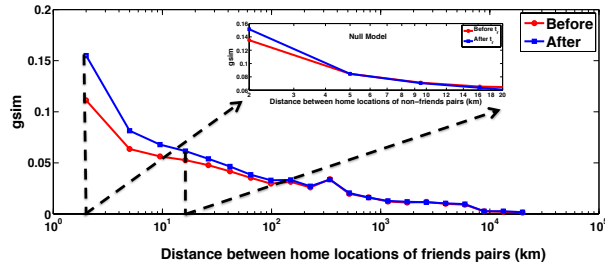


Figure 4: Global influence can possibly explain only up to 2.32% of the global similarity between friends.

the various tests we will use in our analysis⁴.

Knowing the time of friendship formation between Bob and Alice enables us to calculate a similarity value for Bob and Alice before and after becoming friends. A similarity increase *might* be a signal for peer influence. Using the cosine similarity metric, the global similarity between users u and v prior to becoming friends is:

$$gsim_{u,v}^b = \frac{\mathbf{c}_u^b \cdot \mathbf{c}_v^b}{\|\mathbf{c}_u^b\|_2 \|\mathbf{c}_v^b\|_2} \quad (2)$$

where, \mathbf{c}_u^b and \mathbf{c}_v^b are defined analogously to Section 3, as vectors describing the venues that u and v checked-in to before they became friends. In particular, the i^{th} element of \mathbf{c}_u^b , is equal to the number of check-ins that u had in venue i , prior to becoming friends with v . Since $gsim_{u,v}^b$ is computed over the check-ins that took place before u and v got connected, it can be thought as an *inherent* global similarity between these two users.

Once u and v got associated, we can compute a new similarity value as follows:

$$gsim_{u,v}^a = \frac{\mathbf{c}_u^a \cdot \mathbf{c}_v^a}{\|\mathbf{c}_u^a\|_2 \|\mathbf{c}_v^a\|_2} \quad (3)$$

where now vectors \mathbf{c}_u^a and \mathbf{c}_v^a are created as above but using the whole check-in histories of u and v respectively (i.e., both before and after they became friends).

Using our data, we can compare the global similarity between a pair of users before and after becoming friends. Figure 4 presents our results. The value on the x-axis is the distance between the home locations of friends pairs, and the y-axis is the corresponding average global similarity of the pairs. We use logarithmic binning for the home location distance to reduce - to the extent possible - the noise due to fewer samples at the right end of the x-axis. As we can observe, the global similarity does not change significantly after the friendship creation when the home locations of the friends are more than 10km apart. However, for smaller home distances, there is a non-negligible increase in the $gsim$ between the friendship pairs formed. Furthermore, it is worth noting that global similarity values reduce with an increase in home location distance, as one might have expected (the more distant the homes of two friends the less possible is for them to check-in to common venues).

If we further consider the area under the “blue” line to represent the overall global similarity between friends, we can see that on average 88.68% of it can be explained from the *inherent* similarity between pairs of users (as captured by the area under the “red”

⁴Furthermore, the methodology presented in this section itself could be applicable in different settings and hence, beneficial to other researchers.

curve), while the rest 11.32% can be attributed to peer influence. However, we would like to emphasize here that this number serves only as an upper bound for the amount of global peer influence. With the above statistical test we can only quantify the contribution of the inherent similarity between pairs of users on the overall global similarity. Nevertheless, there can be other reasons that can explain the additional 11.32% in the global similarity.

One possible reason for the increase in the global similarity, especially for friends whose home locations are nearby, is the fact that as people accumulate more activity and visit more places, their similarity to other users (friends or not) can increase just by the chance of visiting the same locations. In other words, (the inherent similarity) $gsim_{u,v}$ might be an increasing function with time, regardless of whether u and v form a tie or not. To examine such a possibility, we consider pairs of users (w, z) that have not formed a social tie during the period that our dataset spans. We will pick a reference time t_r at random, and compute their global similarity prior and after t_r . We are especially interested in pairs whose home location distance is less than 20km. Our results from this *null* model are overlayed and zoomed in, within Figure 4. As one can observe, even for pairs of non-friends their global similarity increases with time. If we further calculate the areas under the curves, we can see that approximately 9% of the change in the global similarity of a pair of users after becoming friends can be explained by its *natural* increase with time⁵.

Factoring in the above percentile temporal increase of global similarity, we re-calculate the upper bound on global peer influence, and eventually only at most 2.32% of the global similarity can be attributed to global peer influence. Hence, it appears that **there is no global influence between friends**.

Contextual dependencies: In the above we have considered the check-ins of users at all possible types of venues. However, influence can clearly be context dependent; while in aggregate there is no (or very small) global influence among friends, it is possible that global peer influence exists for certain types of places. For instance, while our friends’ visits at restaurants might not affect us because we have our own taste in food, the same might not be true for nightlife spots. More general, people might have an impact on friends’ decisions about specific types of venues.

To quantify any context dependencies on global influence we perform the same statistical test as above, but instead of considering check-ins to all the venues in vectors \mathbf{c}_u , we only consider check-ins to venues of the specific category under examination. Figure 5 presents our results for five representative, distinct categories of spots in Gowalla. In particular, we consider “Coffee Shops”, “Food” joints, “Pubs”, “Shopping” venues, “Airports”(results for the rest 278 categories are omitted but they do not differ significantly). The left column of figures are the results for the pair of friends, while in the right column are the corresponding results for the null model as above. Table 2 further presents the upper bound on the percentage of global similarity between friends that can be assigned to the global peer influence for the different types of venues. Note that in some cases we obtain negative values for the upper bound of global peer influence. This is essentially an artifact of the time t_r picked for the null model. Nevertheless, even with the most accurate choice of t_r , the upper bound on the global peer influence is not expected to be a much larger positive number⁶. Hence, even if we consider types of places in isolation, there

⁵We would like to emphasize on the fact that this result is only approximate, since an exact result (i) depends on the accurate choice of t_r and (ii) would require the estimation of a function $gsim_{u,v}(t)$. Both are beyond the scope of our work.

⁶We have examined a variety of other strategies for choosing t_r ,

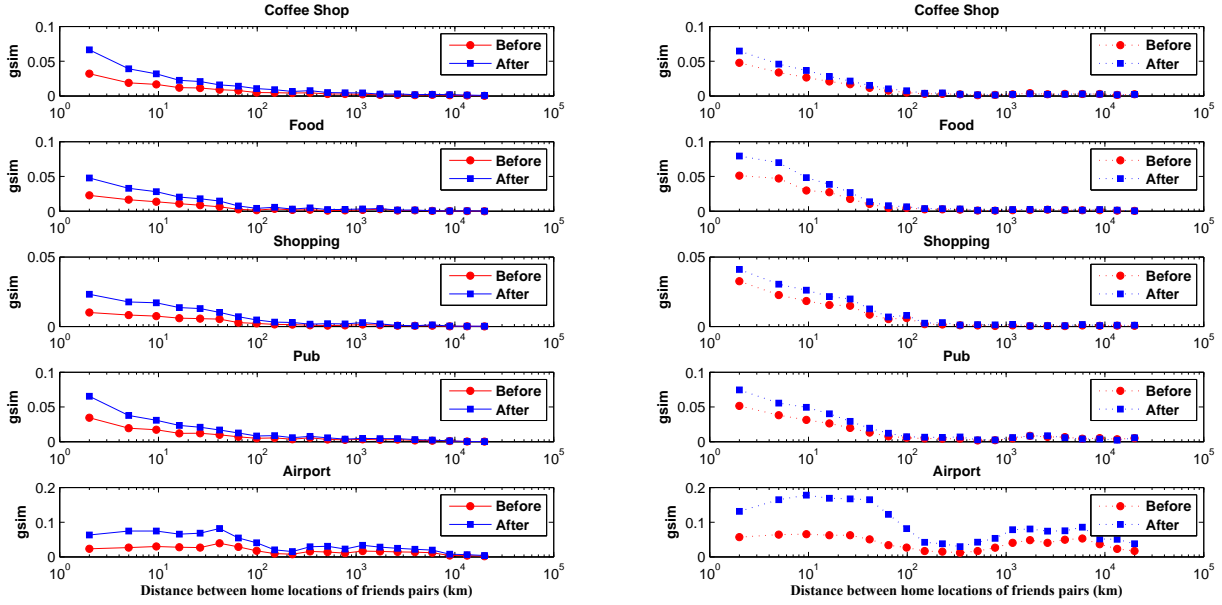


Figure 5: Levels of global peer influence are very small regardless of the venue context.

appears to be no global peer influence on average.

Table 2: Even after considering specific context (i.e., type of places), there appears to be no global peer influence.

Venue type	Upper bound on global influence
Coffee Shop	2.08%
Food	1.05%
Shopping	-4.60%
Pub	-3.13%
Airport	0.04%

4.2 Local Influence

Our previous results support - as one might have expected - the absence of global influence between pairs of friends. However, peer influence mechanism might operate in smaller spatial scales, and hence we seek to examine in this section the existence of a localized version of peer influence. In other words, while Jack is not influenced by his friend Jill (who possibly lives more than 20kms away) at a global scale, he might be influenced when he is around Jill's home location. In order to examine whether there exists **local influence** or not using the above procedure, we would need pairs of friends who had been area co-located (see Section 2.2) in each others home location both prior and after they become friends. This would allow us to estimate both an inherent local similarity as well as a local similarity after becoming friends. However, there are not many such pairs to yield statistically significant results. Hence, we devise a different test.

In particular, we consider pairs of friends, (u, v) , who have been area co-located at u 's and/or v 's home location after becoming friends. In addition, we filter out pairs (u, v) that have been check-in co-located before becoming friends (however, they can have been area co-located). The reason for the latter is to remove from our test-set users that have non-zero inherent similarity and essentially and they all give values close to zero.

to rule out one possible reason for the observed (if any) local similarity. Note here that, the test we will use in what follows cannot account for that. The above filters finally give us 43,618 pairs of friends.

Using these pairs we calculate the local similarity between u and v as follows:

$$lsim_{u,v}^{data} = \frac{\mathbf{c}_u^v \cdot \mathbf{c}_v^v}{\|\mathbf{c}_u^v\|_2 \|\mathbf{c}_v^v\|_2} \quad (4)$$

where, \mathbf{c}_u^v and \mathbf{c}_v^v are as in Equations (2) or (3) but now we are considering only the check-ins of u and v respectively, in venues near v 's home location (i.e., within a radius of 25km from v 's home location).

To reiterate, given our setup, this value of local similarity (large or small) cannot be attributed at any part to inherent similarity between the users, since the specific pairs we examine were not check-in co-located prior to becoming friends (i.e., their similarity - global or local - was zero)⁷. However, we need to compare this local similarity with some benchmark values that capture other possible reasons that lead to the observed behavior. In particular, we devise two reference models that aim in capturing (i) the expected local similarity if u was checking-in at random (Random Reference Model - RRM), and (ii) the expected local similarity if u was choosing his check-ins based on the popularity of the venues (Popularity-based Reference Model - PRM). In both models we retain the structure of the real check-ins, that is, the total number and their categories.

Simply put, let us assume that u has visited v 's home location, and he performed z number of check-ins ten of which where in coffee shops and the rest in restaurants. For our RRM, we will uniformly at random sample ten times all the coffee shops in v 's home location, and $z - 10$ times the local restaurants. This process will generate a synthetic dataset for the check-ins of u in v 's home lo-

⁷Of course, as we also further explain in Section 6, there can be missing "check-in co-locations" of users, not captured from the dataset, due to the voluntary fashion of location sharing in Gowalla.

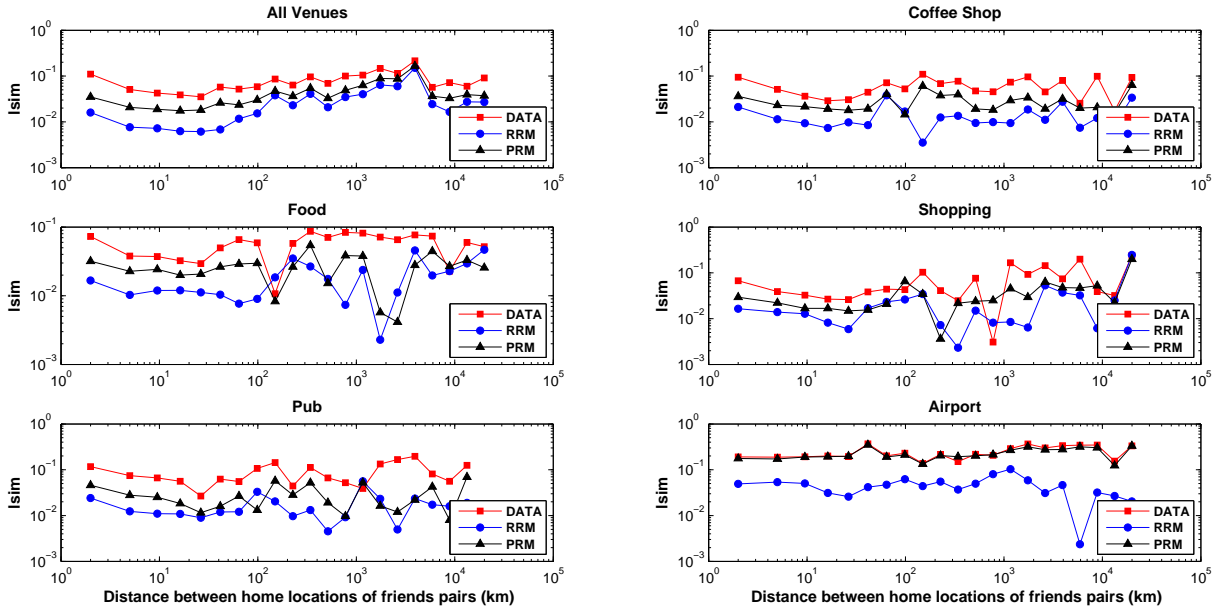


Figure 6: Local similarity as obtained through data and two randomized reference models.

cation, and consequently will give a corresponding vector $\mathbf{c}_{u,RRM}^v$. Similarly, for the PRM, we will follow exactly the same process, but instead of picking venues uniformly at random, we will bias the sampling probabilities based on the popularity of each venue π as captured from the total number of users that have checked-in at π . This will further give us another vector $\mathbf{c}_{u,PRM}^v$.

Using our reference models' vectors for user u we can obtain the reference local similarities between u and v as:

$$lsim_{u,v}^{RRM} = \frac{\mathbf{c}_{u,RRM}^v \cdot \mathbf{c}_v^v}{\|\mathbf{c}_{u,RRM}^v\|_2 \|\mathbf{c}_v^v\|_2} \quad (5)$$

$$lsim_{u,v}^{PRM} = \frac{\mathbf{c}_{u,PRM}^v \cdot \mathbf{c}_v^v}{\|\mathbf{c}_{u,PRM}^v\|_2 \|\mathbf{c}_v^v\|_2} \quad (6)$$

We want to emphasize on the fact that v 's check-in vector in Equations (5) and (6), is obtained from the real data. Furthermore, we run RRM and PRM 100 times for each pair of users and obtain the average reference local similarity.

Figure 6 presents our results. As with global similarity, we present the results obtained both by considering all the check-ins (top left subplot) as well as considering check-ins to specific types of locations (rest of the subplots). As we can see there is some level of local similarity that would have been expected even if people were checking-in completely at random. The percentage of local similarity that can be explained by PRM is even higher, and many times it appears to be the main reason for the levels of local similarity. For instance, for "Airports", the curve obtained from the real data is almost on top of the curve for PRM. Each city typically only has a few airports, among of which, even less support many connections, and therefore, being popular. People will pick these airports not because they are influenced by their peers, but because they are more convenient.

Table 3 summarizes the progressive percentage of local similarity between friends that can be explained by the two reference models (RRM and PRM), as well as the *maximum* possible effect of local influence. As we can see RRM can explain approximately 44% of the observed local similarity when considering all the check-ins, while an additional 16% can be attributed to PRM. Consequently, local peer influence mechanisms can explain up to almost 40% of

the local similarity between friends. Hence, *friends appear to be influenced more easily when they are in proximity as compared to a global scale*. Moreover, the levels of local influence are context dependent. For instance, there appears to be no local peer influence in "Airports" (upper bound of local peer influence is only 10%), but significant levels are observed in "Pubs" (approximately 64%).

Our previous results support the existence of *local* influence between friends. Nevertheless, we have explicitly focused on the activities of pairs of friends around each other's home location. Now, we seek to examine the existence of peer influence in a *third location*. In particular, Jack and Jill can have been area co-located in a third location, not specifically at one of their home locations. We consider a third location, L_{3rd} , as a distant area (25kms far away) from both of the pairs' home locations. For example, Jack's home location is New York City and his friend's Jill's is San Francisco. If they have been area co-located in a third city (e.g., Boston), one might still influence the other. Note here that these situations are included in the global influence study. Nevertheless, they might have been *lost* in the aggregation of all different locations around the globe that users have visited. Hence, we examine them separately in what follows.

For every pair of friends (u, v) we use their combined check-ins to locations different than their home locations. We then exert the DBSCAN clustering method on these check-ins (features are the latitude/longitude pairs). If a cluster identified by the algorithm includes check-ins from both u and v , then we say this pair has been area co-located at L_{3rd} . Furthermore, L_{3rd} is considered to be the centroid of all the check-ins in the specific cluster. Similar to the above local similarity analysis, we use these area co-located pairs in a third location and we calculate a similarity score (which we refer to as remote similarity - *rsim*) as follows:

$$rsim_{u,v}^{data} = \frac{\mathbf{c}_{u,L_{3rd}}^{L_{3rd}} \cdot \mathbf{c}_{v,L_{3rd}}^{L_{3rd}}}{\|\mathbf{c}_{u,L_{3rd}}^{L_{3rd}}\|_2 \|\mathbf{c}_{v,L_{3rd}}^{L_{3rd}}\|_2} \quad (7)$$

where, $\mathbf{c}_{u,L_{3rd}}^{L_{3rd}}$ and $\mathbf{c}_{v,L_{3rd}}^{L_{3rd}}$ are as in Equation (4) but now we are considering only the check-ins of u and v respectively, in venues *near* L_{3rd} (i.e., within a radius of 25km from L_{3rd}). Similarly, we consider two reference models as in Equations (5) and (6):

Table 3: Progressive percentage of local similarity that can be attributed to RRM , PRM and local peer influence.

Venue type	% Explained by RRM	Additional % explained by PRM	Upper bound on local peer influence
All Venues	43.93%	16.29%	39.78%
Coffee Shop	26.32%	21.47%	52.21%
Food	56.02%	8.63%	35.35%
Shopping	47.14%	1.31%	51.55%
Pub	12.55%	23.63%	63.82%
Airport	6.84%	82.94%	10.22%

$$rsim_{u,v}^{\text{RRM}} = \frac{\mathbf{c}_{u,\text{RRM}}^{L_{3rd}} \cdot \mathbf{c}_v^{L_{3rd}}}{\|\mathbf{c}_{u,\text{RRM}}^{L_{3rd}}\|_2 \|\mathbf{c}_v^{L_{3rd}}\|_2} \quad (8)$$

$$rsim_{u,v}^{\text{PRM}} = \frac{\mathbf{c}_{u,\text{PRM}}^{L_{3rd}} \cdot \mathbf{c}_v^{L_{3rd}}}{\|\mathbf{c}_{u,\text{PRM}}^{L_{3rd}}\|_2 \|\mathbf{c}_v^{L_{3rd}}\|_2} \quad (9)$$

where, $\mathbf{c}_{u,\text{RRM}}^{L_{3rd}}$ and $\mathbf{c}_{u,\text{PRM}}^{L_{3rd}}$ are randomly generated by considering venues around the center of the L_{3rd} . In the case of remote similarity, user v is the user that checked-in first in a venue around location L_{3rd} .

Figure 7 and Table 4 present our results, where we can see peer influence also exists in these so-called third locations. Compared to the local influence, the upper bound of similarity explained by peer influence is smaller. Nevertheless, this can be flipped when considering specific context (i.e., categories of venues). To sum up, even though global peer influence does not appear to be significant, if we focus our attention to remote geographic areas that both friends have visited - not necessarily their home locations - peer influence can possibly explain a large part of their similarity.

5. SOCIAL SELECTION

As alluded to Section 1, social selection works between people that have high levels of similarity and can cause the creation of friendships. We further saw in the previous section that users exhibit some inherent similarity, which also increases with time. Also by observing the absolute global similarity values of the null model in Section 4.1, we find that pairs of non-friends exhibit significant levels of global similarity as well. Why then social selection works with specific pairs and not with others?

When examining similar questions, we need to be cautious and in particular to avoid confusing *actual* similarity with what we refer to as “trivial” - or expected - similarity in this study. For instance, there are places that most of the people living in a city will visit, e.g., subway station(s), city hall etc. Such places introduce *trivial* similarity and it does not necessarily mean that the social selection mechanism will be triggered and these people will form social ties. In order to avoid confusions, we would like to emphasize here that trivial similarity is also part of the inherent similarity between two people⁸. However, it does not add valuable information.

In this section we seek to answer the question posed above and investigate the dynamics of the social selection mechanism. In particular, we examine the (network) characteristics of the venues - see Figure 2 - that appear to trigger the selection mechanism. More specifically, we consider pairs of friends that were check-in co-located prior to becoming friends, and hence they exhibited a non-zero level of similarity. In total, we have 84,460 such pairs. For these pairs of friends, we analyze the categories of the places that

they were co-located prior to becoming friends, the *degree* - i.e., number of check-ins - of these places, their clustering coefficient as well as their entropy (to be defined later). As we will see in what follows, all of the above metrics exhibit the same properties for the friends pairs considered.

However, the above results alone are not conclusive. In particular, the common locations of other pairs of users that have been check-in co-located but never formed social ties, can also exhibit the same features. Hence, in order to avoid this sampling bias and to be able to draw safe conclusions, we need a *reference* group for comparison. We randomly pick 84,460 pairs of users that have been check-in co-located (hence, having some non-zero levels of similarity), but never became friends, and we calculate the same statistics for their common locations. Note here that, our random sampling retains in the reference group the same distribution of the home location distances as that for the check-in co-located pairs that eventually became friends. In particular, if there are μ pairs of friends with home distances in the range $[\chi_1, \chi_2]$, we sample uniformly at random μ reference pairs with home distances in the same range.

To preview our results, the features of the common venues of the reference group exhibit significant differences as compared with those of the friends pairs. Furthermore, the common venues of the reference group pairs manifest characteristics of locations that introduce trivial - or expected - similarity (e.g., high degree, low clustering coefficient, large entropy)! These results clearly indicate that **the (online) social selection mechanism works upon non-trivial similarity and can be stimulated by venues with specific features**. In other words, people tend to generate social ties with their peers with whom they exhibit non-trivial/unexpected similarity.

In what follows we introduce the venue metrics we examine and present the details of our results (Appendix B includes statistical significance results for the conclusions).

Venue Category: As mentioned in Section 2.2 every spot in Gowalla is labeled with a category depending on the type of place, and there are 283 possible categories. Hence, we compute the category probability mass function of the common locations for the pairs in the two groups (friends and reference). Figure 8 depicts our results. As we can see, for pairs of users in the reference group, the mass function exhibits a clear *4-modal* distribution. On the contrary, the corresponding mass function for the pairs of friends is closer to a uniform distribution. The four categories-modes of the reference mass function are: “Convention center”, “Interactive”, “Airport” and “Travel/Lodging”. As we can see these are types of places that people can co-locate at, not necessarily because of their similarity or common interests. For instance, there are many reasons that people go to a convention center. Airports are also potentially visited by all people (at the minimum all people that travel). On the contrary, friends tend to co-locate to a variety of places with fairly equal probabilities. Nevertheless, the top-4 places for friends are: “Corporate office”, “Pub”, “Food” and “Coffee shop”. Corporate office is mainly visited by people that work there every day and

⁸ Actually, it might be the case that the “trivial” part of the inherent similarity is the major component that varies with time. Nevertheless, further examining this is beyond the scope of our work.

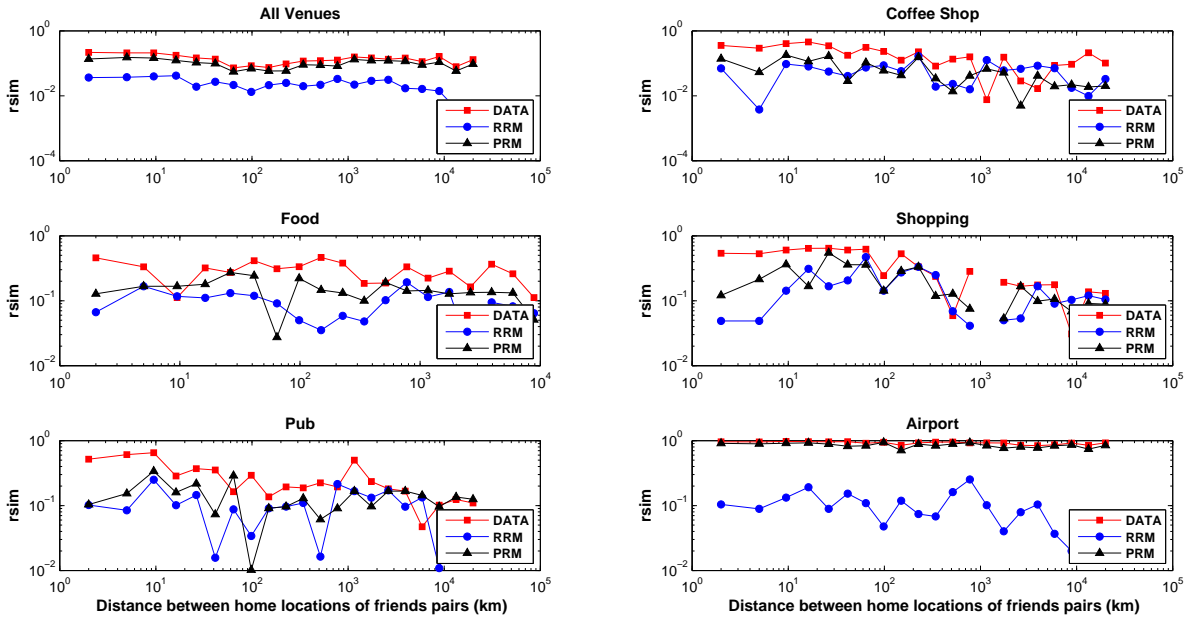


Figure 7: Similarity in a third location as obtained through data and two randomized reference models.

Table 4: Progressive percentage of similarity in a third location that can be attributed to RRM, PRM and local peer influence.

Venue type	% Explained by RRM	Additional % explained by PRM	Upper bound on remote peer influence
All Venues	9.73%	64.9%	25.37%
Coffee Shop	12.94%	0.51%	86.55%
Food	17.62%	9.04%	73.33%
Shopping	38.78%	19.63%	41.58%
Pub	46.49%	20.72%	32.8%
Airport	5.76%	85.77%	8.47%

hence they create tight bonds. Of course “Pub”, “Food” and “Coffee shop” locations can also attract a diverse crowd, especially if these places are popular. Hence, while there is a clear difference at the category distributions between friends and reference pairs, we cannot claim that these results are absolutely conclusive. We will now focus on network characteristics of the venues, which are not tied to the category of the place. Such metrics can possibly make further distinctions even between places of the same category (e.g., two restaurants) and therefore, lead to stronger conclusions.

Venue Degree: Next, we examine the degree of the common venues. In particular, we define as the degree $deg(\pi)$ of a place/venue π , the number of total check-ins in this place. Figure 9 presents our results. As we can see the pair of users that eventually become friends have co-located prior to that to venues with lower average degree as compared to that of the common venues for the pairs in the reference group.

Venue Clustering Coefficient: If there are n_π unique people that have checked-in at place π , we define the clustering coefficient of π as follows:

$$CC_\pi = \frac{k}{n_\pi(n_\pi - 1)/2} \quad (10)$$

where k is the number of friendship pairs between the n_π users that have checked-in at π . Equation 10 is essentially the direct extension of the definition of the local clustering coefficient of a graph in our context. A high clustering coefficient for a venue translates

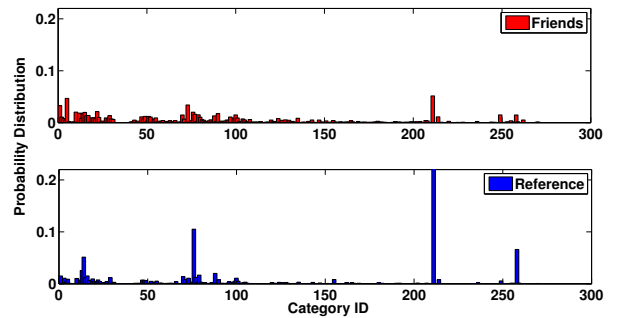


Figure 8: Friend and non-friend pairs have only 4 categories in common in their top 10 categories.

to a location where people who visit it form a tightly connected social group. Figure 10 depicts our results, and as we can see pairs who become friends tend to co-locate to venues with higher CC as compared to the common places of the non-friends pairs of our reference group.

Entropy: Cranshaw *et al.* [7] use the notion of entropy of a location as a measure of its diversity. In particular, if $P_\pi(u)$ is the fraction of check-ins at place π contributed by user u , then the

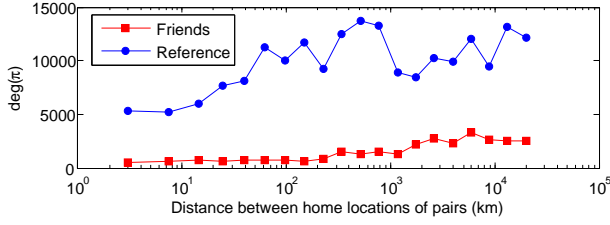


Figure 9: Users that form social ties co-locate to venues with low degree.

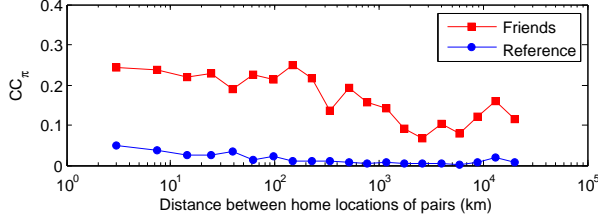


Figure 10: Venues with higher CC are more likely to form friendships.

entropy of π is given by:

$$e(\pi) = - \sum_{u: u \in S} P_{\pi}(u) \log(P_{\pi}(u)) \quad (11)$$

where S is the set of users that have checked-in venue π .

From the definition of $e(\pi)$ we can see that when a place is visited by many people in fairly equal (and hence, small) proportions, its entropy will be high. Simply put, high entropy corresponds to places such as transportation hubs and malls that exhibit large diversity with regards to people they “attract”. On the other hand, when the mass of $P_{\pi}(u)$ is concentrated only to a few people, the diversity in this location is small and so is the entropy.

In Figure 11 we present the average entropy of the common locations of the pairs in our reference, non-friends group, and that of the friends pairs prior to forming their tie. As we can see pairs who become friends after being co-located, have co-located to venues with lower average entropy compared to the average (common venues) entropy of our control group pairs.

Note here that places with high degree, low clustering coefficient and high entropy are essentially places that are responsible for trivial similarity. These are locations with large (high $deg(\pi)$), diverse (high $e(\pi)$) crowds that are disconnected in the social plane (low CC_{π}). These are places that people visit not essentially because they are in their preference but because they have to (e.g., airports, train stations, big university campus, medical centers etc.). On the other hand places with low degree, high clustering coefficient and low entropy, profile venues where people that visit them know each other and are actually similar, since they do not attract a large diverse public. Hence, based on our results we can clearly see that in order for the social selection mechanism to be triggered, actual similarity between users is required. Trivial similarity caused by co-locations in places with high degree, low clustering coefficient and high entropy, is not enough.

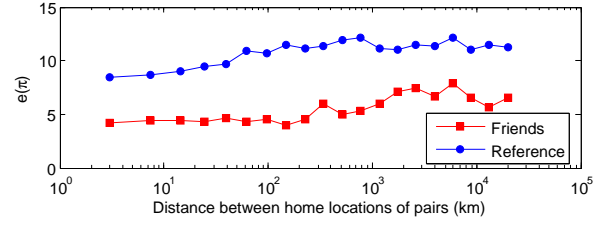


Figure 11: Users that form social ties co-locate to venues with lower average entropy compared to the reference pairs.

6. DISCUSSION AND CONCLUSIONS

In this work, we examine the peer influence and social selection mechanisms in the context of locations visited by friends using a longitudinal dataset from a location-based social network. We find that strong evidence for peer influence exist as long as friends are in proximity, and it is context dependent. In particular, for specific types of places (e.g., nightlife spots) users can influence their peers more as compared to other types of locations (e.g., airports). We also reveal, that social selection works upon non-trivial similarity and there are particular venues - with specific network characteristics - that can trigger the social selection mechanism.

We acknowledge that our study is limited by the information available in the dataset used. For instance, while we know the friendship creation time on the system between a pair of friends, we use the implicit assumption that this is also the actual time of the real-world friendship formation. Of course, this might not be always true. Furthermore, our dataset can exhibit biases with regards to the demographics of people that are using systems like Gowalla. Our results inevitably do not extensively take into consideration the behavior of parts of the population that are possibly underrepresented in the dataset (e.g., older people that might not be as technology savvy). Finally, the voluntary nature of location sharing can possibly introduce another type of bias for our analysis. More specifically, the activities shared in Gowalla (or any other similar social media platform - e.g., Foursquare etc.) are only partially reflecting people’s trajectory. Nevertheless, these are essentially limitations shared - partially or entirely - by any study that is based on digital trails of human activities.

Despite the above limitations, we believe that our findings can stimulate further research on the topic and will contribute to eventually obtaining a more clear understanding on how people create social ties and act in real space. This understanding can facilitate a variety of applications. For example, it can drive enhancements in socially-aware recommender systems or assist venue managers identify potential users for targeted advertisement. In the future, we seek to identify methods for controlling to the extent possible for the above biases and exactly quantify the time varying global similarity between people and decompose it to its various parts (e.g., trivial inherent, actual inherent etc.). We also opt to examine groups of friends (rather than only pairs as in our current work) and the ways peer influence operates in such settings.

7. ACKNOWLEDGMENTS

We would like to thank Prof. Cecilia Mascolo, Dr. Salvatore Scellato and Anastasios Noulas for providing us with the Gowalla longitudinal dataset. We would also like to thank the anonymous reviewers for their valuable comments that helped us improve our work.

8. REFERENCES

- [1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *ACM KDD*, 2008.
- [2] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.
- [3] E. Bakshy, J. Hofman, W. Mason, and D. Watts. Everyone’s an influencer: quantifying influence on twitter. In *ACM WSDM*, 2011.
- [4] E. Bakshy, B. Karrer, and L. Adamic. Social influence and the diffusion of user-created content. In *ACM EC*, 2009.
- [5] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *ACM KDD*, 2011.
- [6] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *ACM KDD*, 2008.
- [7] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *UBICOMP*, 2010.
- [8] P. Erdős and A. Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5: 17-61, 1960.
- [9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *ACM KDD*, 1996.
- [10] S. Goel, D. Watts, and D. Goldstein. The structure of online diffusion networks. In *ACM EC*, 2012.
- [11] G. Kossinets and D. Watts. Empirical analysis of an evolving social network. In *Science* 311:88-90, pages 405–450, 2006.
- [12] G. Kossinets and D. Watts. Origins of homophily in an evolving social network. In *American Journal of Sociology*, Volume 115, Number 2, pages 405–450, 2009.
- [13] K. Lewis, M. Gonzalez, and J. Kaufman. Social selection and peer influence in an online social network. In *Proceedings of the National Academy of Sciences*, Volume 109, Number 1, pages 405–450, 2012.
- [14] M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. In *Annual Review of Sociology*, 27:415-44, 2001.
- [15] S. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *ACM KDD*, 2012.
- [16] M. Newman. Mixing patterns in networks. In *arXiv:cond-mat/0209450v2 [cond-mat.stat-mech]*, 2002.
- [17] K. Pelechrinis and P. Krishnamurthy. Location affiliation networks: Bonding social and spatial information. *ECML/PKDD*, 2012.
- [18] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. In *AAAI ICWSM*, 2011.
- [19] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *ACM KDD*, 2011.
- [20] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *ACM KDD*, 2009.
- [21] J. Tang, S. Wu, and J. Sun. Confluence: Conformity influence in large social networks. In *ACM KDD*, 2013.
- [22] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A. Barabási. Human mobility, social ties, and link prediction. In *ACM KDD*, 2011.

APPENDIX

A. ERDŐS-RÉNYI RANDOM GRAPHS

In a random graph model some properties of the network are fixed, while others are generated randomly. In the Erdős-Rényi random graph model, denoted as $G(n, m)$, we fix the number of nodes to n and the number of edges to m . $G(n, m)$ is then a probability distribution $P(G)$ over all the possible networks G such that if G has n vertices and m edges and Ψ is the number of such (simple) graphs, then $P(G) = 1/\Psi$; otherwise $P(G) = 0$. A slightly different and more tractable model, is the $G(n, p)$. In this case the number of nodes is still fixed (n) but now instead of fixing the actual number of edges in the network, we fix the probability of an edge between any two nodes to be equal to p . More details on random graph models can be found in [8].

B. STATISTICAL SIGNIFICANCE RESULTS

While we observe in Figures (9)-(11) that the average degree, clustering coefficient and entropy of the common venues for the friend’s dataset are different from that of the reference group of user pairs, we further delve into the statistical significance of these results. In particular, every point in Figure 9 corresponds to the mean value for the degree of the common venues of friends residing in distance d , $\mu_{deg}^f(d)$ (bottom line) or of the reference pairs of users $\mu_{deg}^r(d)$ (top line). In order to examine whether the difference observed in the Figure is statistically significant we perform a one-tailed t-test on the mean values for each distance-bin d . In particular, the hypothesis test is:

$$H_0 : \mu_{deg}^f(d) = \mu_{deg}^r(d) \quad (12)$$

$$H_1 : \mu_{deg}^f(d) < \mu_{deg}^r(d) \quad (13)$$

The p-values indicate that for all distances d the null hypothesis can be rejected at the 95% significance level, while for the majority of the cases (and in particular for small d) it can also be rejected at the 99% significance level. Similar results are also obtained for the differences observed at the clustering coefficient and the entropy of the venues. Given that the t-test makes the assumption of normality in the data, we also performed the Mann-Whitney U test for the median. The latter does not have the normality assumption and lead us to similar conclusions with respect to the statistical significance of the differences in the network characteristics for the two sets.