EBM - An Entropy-Based Model to Infer Social Strength from Spatiotemporal Data

Huy Pham
Computer Science Dept.
Univ. of Southern California
huyvpham@usc.edu

Cyrus Shahabi
Computer Science Dept.
Univ. of Southern California
shahabi@usc.edu

Yan Liu
Computer Science Dept.
Univ. of Southern California
yanliu.cs@usc.edu

ABSTRACT

The ubiquity of mobile devices and the popularity of location-based-services have generated, for the first time, rich datasets of people's location information at a very high fidelity. These location datasets can be used to study people's behavior - for example, social studies have shown that people, who are seen together frequently at the same place and at the same time, are most probably socially related. In this paper, we are interested in inferring these social connections by analyzing people's location information, which is useful in a variety of application domains from sales and marketing to intelligence analysis. In particular, we propose an entropy-based model (EBM) that not only infers social connections but also estimates the strength of social connections by analyzing people's co-occurrences in space and time. We examine two independent ways: ***diversity*** and ***weighted frequency***, through which co-occurrences contribute to social strength. In addition, we take the characteristics of each location into consideration in order to compensate for cases where only limited location information is available. We conducted extensive sets of experiments with real-world datasets including both people's location data and their social connections, where we used the latter as the ground-truth to verify the results of applying our approach to the former. We show that our approach outperforms the competitors.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining, Spatial databases and GIS; H.3.3 [**Information Search and Retrieval**]: Retrieval models

Keywords

Social network; social strength; spatiotemporal; geospatial; spatial; data mining; social computing

1. INTRODUCTION

In the past, finding a person's location involved some detective work by following the person clandestinely. However, these days given all the online trails we leave behind, people's locations can be tracked at a very high resolution effortlessly. The location data can be inferred implicitly, for example from one's credit-card transactions or activities on mobile devices (through cell-phone towers, GPS, or WiFi hotspots). It can also be released explicitly, for example when someone distributes geo-tagged contents (e.g., tweets, uploads photos on Instagram, Flickr or Facebook), interacts with location-based-services online (e.g., Foursquare check-ins), or through a mobile app (e.g., Highlight, Glancee). Such a collection of people's locations over time (aka ***spatiotemporal*** data) is a rich source of information for studying various social behaviors. In particular, the one behavior we are interested in this paper is whether social relationships among people can be inferred from such a collection. The intuition is that if two people have been to the same places at the same time (aka ***co-occurrences***), there is a good chance that they are socially related. The ultimate goal is to derive the social-network of people and the social strength from their real-world location data as opposed to (or in addition to) their online activities.

The applications for such a physically inferred social network are plenty - it not only subsumes all the applications of online social networks such as marketing applications (e.g., target advertising, recommendation engines such as friendship suggestions), social studies (e.g., identifying influential people) and cultural studies (e.g., to examine the spreading patterns of new ideas, practices and rumors), but also has its own unique applications. For example, the network can be used to identify the new (or unknown) members of a criminal gang or a terrorist cell or it can be used in epidemiology to study the spread of diseases through human contacts.

However, the problem of inferring social connections from people's spatiotemporal data is particularly challenging for many reasons. First, it is not clear which attributes of co-occurrences should be measured to infer social connection. For example, if the number of co-occurrences of two people, called ***frequency***, is only considered, then one may arrive at a wrong conclusion about their social relationship. To illustrate, suppose two people study at the same library around the same time every day, which results in high frequencies, but they may not even know each other. This erroneous conclusion can be attributed to the fact that the library is a popular location and the observation that two people only co-occur at the library is not a strong indication of social connection. On the other hand, a few co-occurrences in a small private place are perhaps a better indication of friendship. Or alternatively, several co-occurrences at different popular places (e.g., coffeehouses, restaurants) may also be a better indication of friendships. Second, we are interested in inferring more information about social connections such as how close of a relationship two people have (aka ***social strength***). Third, there may be a lot of missing data, as people's location data may be sparse. Fourth, the spatiotemporal data is often extremely large, in the order of gigabytes, which could render

SIGMOD'13, June 22–27, 2013, New York, New York, USA.

the inference algorithms inefficient, taking too much time and/or resources to perform.

Our work is motivated by the following pioneering studies. First, Eagle et al. [9] showed there is a correlation between people's co-occurrences and their social connections by reporting on a study they conducted on a number of students and faculty members from a research institute. Following this observation, Crandall et al. [6] developed a probabilistic model to estimate the probability of two people being friends given their co-occurrences in space and time. Cranshaw et al. [7], on the other hand, introduced various features of co-occurrences and then utilized machine learning techniques to classify pairs of users as friends or not. Distinguished from these approaches is the work done by Li et al. [12], who represented each user's visit pattern as a trajectory, and friendship between two users is determined as the similarity between their trajectories. Lastly, in previous work [15], we studied this problem by introducing two properties, commitment and compatibility, that any social distance measure should follow to correctly infer social connections. While these studies showed the validity, importance and feasibility of inferring social connections from co-occurrences, they either made many simplifying assumptions (see Section 6) or failed to address some of the subtle questions about social connections. This paper is built on these studies by addressing the following questions: 1) inferring the strength of social connections from co-occurrences, 2) avoiding overestimation or underestimation of social connections by taking into account various features of co-occurrences and 3) focusing on the efficiency of the algorithms.

A naive approach to estimate the social strength is to simply count the number of unique locations two people co-occurred as their social strength. However, this measure would consider different locations equally important as it ignores the number of co-occurrences at each location. To address this problem, one could sum up the number of co-occurrences of two people across different locations as a measure of their social strength. The problem with this approach is that it may overestimate coincidences. For example, 10 random encounters at a crowded coffee shop (called coincidences) are considered 10 times more important than 1 interactive meeting at a private office.

To remedy for these shortcomings, we propose an Entropy-Based Model, named **EBM**, which successfully infers social strengths from spatiotemporal data with high accuracy. With EBM, we first use the Shannon entropy to measure the diversity of co-occurrences, which, for each pair of people, uses the number of their co-occurrences at each location to derive a relative co-occurrence measure, and use *only* diversity as social strength. This measure may give higher importance to outliers (i.e., local frequency), and thus may still overestimate the social strength due to coincidences. Hence, we generalize Shannon by using the Renyi entropy that can look at the global pattern of co-occurrences per user pair and has the flexibility of giving more or less weight to outliers by varying the order of diversity q. However, Renyi is still blind to the characteristic of a location (e.g., whether the location is a crowded public coffeehouse vs. a private office). Therefore, we incorporate weighted frequency, which utilizes location entropy to weigh each co-occurrence differently depending on the characteristics (crowdedness) of the location, thus it captures minor co-occurrences that can be a significant indication of a social connection. A summary list of our contributions is provided below.

- EBM quantifies the strength of each social connection by considering how diverse the distribution of the co-occurrences is in the context of locations (aka ***diversity***).
- EBM avoids overestimation by discounting coincidences, utilizing ***diversity's order*** - an important property of diversity.
- EBM compensates for the data sparseness by taking into account the local characteristics of locations (e.g., location popularity).
- We evaluated EBM using a large real-world dataset collected by a location-based social network called Gowalla. We 1) use **only** the Gowalla's location data to infer users' social connections, and 2) use the Gowalla's social-network as the **ground truth**.
- Our evaluation shows that EBM's predicted social strength is consistent with ground truth. 88% of social strengths are correctly predicted by our model.
- As for inferring friendships, by using the the diversity, we achieve a precision of 96.5%, but the recall was low due to the data sparseness. However, after incorporating location entropy into EBM, we improve the recall by a factor of 1.8.
- Our EBM's algorithm is parallelizable, thus can be implemented using the MapReduce framework in order to be efficient for massive data, which is critical in any online applications.
- Finally, we experimentally compared our model to the previous studies, including GEOSO [15], probabilistic model [6], the feature model [7] and the trajectory model [12], and the results show that EBM outperforms them in both efficiency and accuracy.

The rest of this paper is organized as follow: In Section 2, we formally define the problem and discuss the preliminaries necessary to construct the model. In Section 3 we explain the EBM model and how it answers the questions raised by the paper. Section 4 describes the optimization of the algorithm to deal with large datasets. We describe the experiment in Section 5, related work in Section 6, and make the conclusion of the paper in Section 7.

2. PROBLEM DEFINITION

In this section we will give the formal definition of the problem and introduce the notations and preliminaries that are used later to formulate the EBM model.

2.1 The Problem

As a user checks in at a location, the following information will typically be recorded and sent to the server: User's **ID** - u; User's **location** - l, which consists of latitude and longitude's values and a unique ID that represents a specific place such as a shopping center, a theater, a living house, etc; the **time** of the check-in - t. Therefore users' check-ins can be represented as a set of ***user-location-time*** triplets $< u, l, t >$, each of which states that User u visited location l at time t. We give the formal definitions of the problem as follow:

DEFINITION: ***Social strength*** is a quantitative measure that tells how socially close two people are.

DEFINITION: Given a set of users $U = (u_1, u_2, ..., u_M)$, a set of locations $L = (l_1, l_2..., l_N)$ and a set of check-ins in the forms of *user-location-time* triplets $< u, l, t >$, the problem is to infer the social strength for each pair of users.

Note: The only time-related input to the problem is the check-in's time. Another possible input that can greatly influence social strength is the amount of time two users stayed together at a place, often referred to as ***length of stay***. However, most location-based social networks nowadays, such as Facebook, Foursquare, Yelp, etc., *neither* record the length of stay *nor* provide such services. In fact, users check in at places, but never check out. Therefore, we design our model to capture such reality by limiting the input to only User IDs, locations and check-in times. Desirably, the length of stay can be a consideration in our future work once such information becomes readily available.

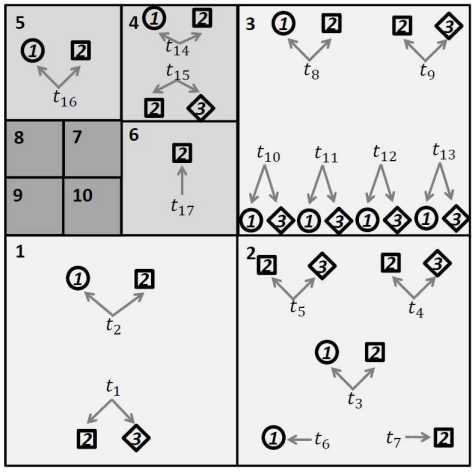


Figure 1: A quadtree storing areas of different levels of popularity, and the visits of Users $1, 2$ and 3

2.2 Preliminaries

2.2.1 Representation of Location

One popular way of storing visited locations is to use a grid to uniformly partition the space into disjoint cells of ***equal size*** [6, 15], where each cell represents only one ***place***, so that any two people, who check in within the same cell at the same time, are considered to have a ***co-occurrence***. However, a uniform grid is inflexible and inefficient due to the two following reasons: **1)** In a crowded area such as a downtown, a place, which represents the location of a co-occurrence, say a shopping center, is often much smaller compared to a place in a sparse mountainous region, say a national park. Hence, the method of partitioning the space into equal cells is not applicable here. **2)** Reducing the size of all the cells to fit small places in crowded areas would result in much waste of storage resources and look-up time in sparse areas, meanwhile, increasing the cells' size to fit to large places in sparse regions would result in misinterpreting the co-occurrences in crowded areas.

Therefore, to efficiently store spatial data, we use quadtree [17]. Fig. 1 shows a quadtree, where each quadrant, called ***cell***, has a unique ID, numbered from 1 to 10. Three users are shown as circles, diamonds and squares uniquely identified with user IDs 1, 2 and 3. The arrows show that a user checked in at the cell at time t_i. The darker, the denser the area. Geo-points inside a cell share the same cell ID, which is used along with time to determine co-occurrences. For example, looking at cell 1, we say Users 1 and 2 co-occurred at cell 1 at time t_2.

For simple presentation, we set the capacity of each cell of the quadtree to 1 so that each cell can cover a maximum of ***one*** place (in the experiment, the capacity is more than 1). The construction of the tree can be done by recursively dividing an area into four equal quadrants until each quadrant holds only one place.

2.2.2 Visit Vector

The visit history of a user is represented by a ***visit vector***, which shows the cell IDs and the check-in time. For example, the visit vector for User 1 in Fig. 1 is $V_1 = (< t_2 >, < t_3, t_6 >, ...)$, which states that User 1 visited cell 1 at time t_2; visited cell 2 at time t_3 and t_6, etc. The general format of the visit vector of User i is:

$$V_i = (< t_{1,1}, ..., t_{1,i_1} >, ..., < t_{M,1}, t_{M,2}, ..., t_{M,i_M} >) \quad (1)$$

where M is the number of leaves in the quadtree.

2.2.3 Co-occurrence Vector

If two users checked in at the same location within a time-interval τ, then we say that they have a co-occurrence. τ is an application-dependent parameter and can be set experimentally. Correspondingly, a ***co-occurrence vector*** between User i and User j represents all the co-occurrences of Users i and j is:

$$C_{ij} = (c_{ij,1}, c_{ij,2}, ..., c_{ij,M}) \quad (2)$$

where $c_{ij,l}$ is the number of co-occurrences between Users i and j at location l, which is referred to as ***local frequency***, which will be used throughout this paper.

For example, given the visit vectors of User 1 and User 2:

$V_1 = (< t_2 >, < t_3, t_6 >, < t_8, t_{10}, t_{11}, t_{12}, t_{13} >, < t_{14} >, < t_{16} >, 0, 0, 0, 0, 0)$

$V_2 = (< t_1, t_2 >, < t_3, t_4, t_5, t_7 >, < t_8, t_9 >, < t_{14}, t_{15} >, < t_{16} >, < t_{17} >, 0, 0, 0, 0)$

we see that the two users have one co-occurrence at location 1 at time t_2, one co-occurrence at location 2 at time t_3, etc, therefore the co-occurrence vector between User 1 and User 2 is:

$$C_{12} = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$$

Usually, a user can only visit a limited number of locations, which makes the visit vector and the co-occurrence vector sparse (containing many zeroes). Therefore, we will introduce an alternative data structure and storage for optimization in Section 4. For now, we use this format to simplify the presentation.

3. THE EBM MODEL

The goal of this section is to devise a model, named ***Entropy-based Model*** (EBM), to quantify ***social strength*** between two users from their co-occurrence vectors. The overview diagram of EBM is shown in Fig. 2. In Section 3.1, we start by utilizing the diversity of the co-occurrence vectors as the main contributing factor to social strength. Consequently, we use the ***Shanon entropy*** to measure the diversity of co-occurrences (see Section 3.2), but we observe that this measure may overestimate the strength of social connections due to the impact of ***coincidences***, which are the case when people happen to co-occur by chance, but do not interact with each other, especially in crowded places such as downtown and shopping centers. Hence, we generalize Shanon entropy to the ***Renyi entropy*** (see Section 3.3), which gives us the flexibility of controlling how much coincidences can contribute to diversity via a parameter q, called the ***order of diversity***. Finally, to compensate for the problem of data sparseness, we incorporate ***weighted frequency***, which in turn uses ***location entropy***, into our model in Section 3.6 to increase the impact of co-occurrences at uncrowded places even at low frequencies to the strength of social connections. The resulting social strength is the ultimate measure that describes how close two people are based on the history of their spatiotemporal information.

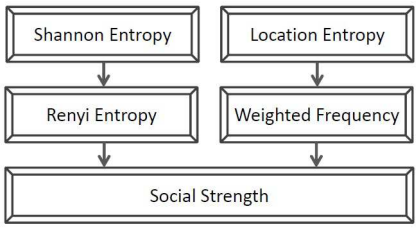


Figure 2: Diagram of EBM - Social strength is formulated via Renyi Entropy and Weighted Frequency

3.1 Diversity in Co-occurrences

The concept of diversity has long been used in Physics, Economics, Ecology, Information Theory, etc, as a quantitative measure to characterize the richness of a system [10, 11, 19]. Specifically, in Ecology, diversity is used to measure how diverse an ecosystem is; in the simplest case, it equals the number of different species in an ensemble. In Statistical Thermodynamics, diversity is the number of micro-states, in which a system can be [18].

Consider the co-occurrence vectors for each user pair in Fig. 1,

$$C_{12} = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$$
$$C_{23} = (1, 2, 1, 1, 0, 0, 0, 0, 0, 0)$$
$$C_{13} = (0, 0, 4, 0, 0, 0, 0, 0, 0, 0)$$

we see that User 1 and User 2 have 5 co-occurrences, so do User 2 and User 3. However, in the former case the co-occurrences are spread over 5 different locations, while in the latter case the co-occurrences happened in 4 different locations. Even simpler is the case of User 1 and User 3, for which all co-occurrences happened in one location - cell 3. We say that C_{12} is *more diverse* than C_{23}, and C_{23} is *more diverse* than C_{13}.

Intuitively, people, who are socially connected, tend to visit *various* places together [5, 6, 7, 9, 15]. This intuition is captured by our model as *how diverse* their co-occurrences are. Applying the general definition of diversity in [20], we formally define *diversity* in our model:

DEFINITION: Diversity is a measure that quantifies how many effective locations the co-occurrences between two people represent, given the mean proportional abundance of the actual locations.

3.2 Formulation of Diversity through Shannon Entropy

In this section, we use Shannon entropy, then we will extend EBM to a more generic one, called Renyi entropy, in Section 3.3.

First, we define the notations and quantities that will be used to construct EBM. $r_{i,j}^{l,t} =< i, j, l, t >$ is a co-occurrence of User i and User j at location l and at time t. $R_{ij}^{l} = \bigcup_t r_{i,j}^{l,t}$ is the set of co-occurrences of User i and User j, which happened at location l. R_{ij} is the set of all co-occurrences of User i and User j at all locations: $R_{ij} = \bigcup_l R_{i,j}^{l} = \bigcup_{l,t} r_{i,j}^{l,t}$

The probability that a randomly picked co-occurrence from the set R_{ij} happened at location l is:

$$P_{ij}^{l} = \frac{|R_{ij}^{l}|}{|R_{ij}|} \tag{3}$$

If we randomly pick a co-occurrence from the set R_{ij} and define its location as a random variable, then the uncertainty associated with this random variable is defined by the Shannon entropy for User i and User j as follow (the upper index S denotes *Shannon*):

$$H_{ij}^{S} = -\sum_{l} P_{ij}^{l} \log P_{ij}^{l} \tag{4}$$

Formulation of diversity: There exists a distinction between entropy and diversity, in which entropy often acts as the index of diversity. For illustration in the former research [11], Jost *et al* compared the *roles* of entropy and diversity as the radius and the volume of a sphere, respectively, where radius is used to calculate volume; and showed their relationship:

$$D = \exp(H) \tag{5}$$

where D is diversity, which indicates how diverse an ensemble is. Following this strategy, we construct the EBM model, where D shows how diverse the co-occurrences of two users are in terms of locations. The diversity of the co-occurrences of User i and User j, which is defined in Equation 5, becomes:

$$D_{ij} = \exp\left(H_{ij}^{S}\right) = \exp\left(-\sum_{l} P_{ij}^{l} \log P_{ij}^{l}\right) \tag{6}$$

Since we already defined the co-occurrence vector in Equation (2), we can rewrite the expression of diversity D_{ij} in terms of the co-occurrence vector as follow:

$$D_{ij} = \exp\left(-\sum_{l, c_{ij,l} \neq 0} \frac{c_{ij,l}}{f_{ij}} \log \frac{c_{ij,l}}{f_{ij}}\right) \tag{7}$$

where $f_{ij} = \sum_l c_{ij,l}$ is the total number of co-occurrences of User i and User j, named *frequency*. Note the difference between *frequency* f_{ij} and *local frequency* $c_{ij,l}$; the *frequency* of two users is the sum of all their *local frequencies*. From Equations (5), (6) and (7), we have observations:

- The higher the number of co-occurrence locations, the higher the uncertainty given by the Shannon entropy, and consequently the higher the diversity.
- If the number of co-occurrence locations is fixed, the diversity and the Shannon entropy reach their maximums when all the probabilities in Equations (5) and (6) are equal to each other.

To demonstrate the observations, let us consider the example of a group of three users in Fig. 1. The co-occurrence vectors, Shannon Entropy, Diversity value, the diverse information, the likelihood of coincidences and the probability of being friends for each pair of users are summarized in Table 1.

From Table 1, we see that C_{12} has the highest value of diversity due to the spread of co-occurrences over more locations, or in other words, C_{12} is the most diverse, followed by C_{23}, then followed by C_{13}, which is the least diverse as all the co-occurrences happened at only one place - cell 3. For C_{12}, the numbers of all co-occurrences are equal to each other (in the first five cells), which produces the maximum value of Shannon entropy, and consequently, the maximum value of diversity among the three co-occurrence vectors, together with the highest number of co-locations (i.e., 5), which makes it the most diverse, therefore, suggesting a high probability of User 1 and User 2 being friends. Furthermore, the value of diversity D_{12} coincides with the number of co-occurrence locations, which is the reason why diversity is often referred to as the ***effective number of states*** (in Statistical Mechanics [18]) or ***effective number of species*** (in Ecology [11]).

In addition, we see that the diversity of C_{23} (3.789) is less than the number of co-occurrence locations (i.e., 4), which is due to the fact that it has two co-occurrences at one place - cell 2 (*less diverse*), as compared to the case of C_{12}, where all co-occurrences are uniformly spread over five different cells (*more diverse*).

We also observe an interesting point where all four co-occurrences of User 1 and User 3 happened at a popular place - cell 3, which has been visited by all three users and has the highest number of co-occurrences in total. This fact implies the high likelihood of coincidences between User 1 and User 3, for example, they might just happen to be at a crowded public place (such as a shopping center or a library) at the same time. Therefore, even four co-occurrences in such a crowded place might still say very little about a possible social connection. We have the following observation:

Observation: A high number of co-occurrences at *only one place* might be an indicator of a friendship if the place is unpopular and uncrowded, but they might suggest the likelihood of coincidences in popular and crowded places. Shannon Entropy and its corresponding diversity, however, would treat multiple co-occurrences as coincidences independent of where took place. Also, Shannon Entropy does not allow us to adjust the impact of this type of co-occurrences on the diversity. Therefore, it is necessary to examine

Table 1: Example of Diversities

Co-occurrence Vector	Shannon Entropy	D_{ij} Value	Diversity	Likelihood of Coincidences	Prob. of a Friendship
$C_{12} = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$	1.609	5.000	High	Low	High
$C_{23} = (1, 2, 1, 1, 0, 0, 0, 0, 0, 0)$	1.332	3.789	Medium	Medium	Medium
$C_{13} = (0, 0, 4, 0, 0, 0, 0, 0, 0, 0)$	0.000	1.000	Low	High	Low

these issues to avoid any false predictions that coincidences might cause. We will address this in Sections 3.4 and 3.5.

3.3 Renyi Entropy-based Diversity

As we discussed in Section 3.2, even though the Shannon Entropy (and its corresponding diversity) can capture the likelihood of a social connection between two people, it cannot distinguish the cases when coincidences *might* or *might not* happen and does not allow us to adjust the impact of coincidences. Rennyi entropy and its corresponding diversity, on the other hand, will give us the utility to control how much coincidences can contribute to diversity. In fact, Shannon entropy is just a special case of Renyi entropy.

Consider the general case of entropy - Renyi entropy, given as:

$$H_{ij}^R = \left(-\log \sum_l \left(P_{ij}^l \right)^q \right) / (q-1) \tag{8}$$

where $q \geq 0$ is the order of diversity. The diversity given by Equation 5 becomes (The upper index R denotes *Renyi*):

$$D_{ij} = \exp(H_{ij}^R) = \exp\left[\left(-\log \sum_l \left(P_{ij}^l \right)^q \right) / (q-1) \right]$$

$$= \left[\exp\left(\log \sum_l \left(P_{ij}^l \right)^q \right) \right]^{1/(1-q)} = \left[\sum_l \left(P_{ij}^l \right)^q \right]^{1/(1-q)} \tag{9}$$

$$= \left[\sum_{l, c_{ij,l} \neq 0} \left(\frac{c_{ij,l}}{f_{ij}} \right)^q \right]^{1/(1-q)} \tag{10}$$

Equation (10) expresses the diversity in terms of a co-occurrence vector. The elegance of using the Renyi entropy in our problem lies inside the parameter q, called the ***order of diversity***, which indicates its ***sensitivity*** to the local frequency $c_{ij,l}$ [16]. Specifically:

- When $q > 1$ the Renyi entropy H_{ij}^R, and consequently the diversity D_{ij}, more favorably considers the high values of $c_{ij,l}$, which are the more popular events. In other words, the higher the local frequency $c_{ij,l}$, the more weight it gets from the diversity or the more impact the local frequency can make on diversity.
- When $q < 1$, in opposite, the diversity tends to give more weight to the local frequencies with low-values $c_{ij,l}$.
- When $q = 0$, the diversity is completely ***insensitive*** to $c_{ij,l}$ and gives the pure number of co-occurrence locations.
- Case $q = 1$: The Renyi entropy favors local frequencies $c_{ij,l}$ in opposite ways when $q < 1$ versus when $q > 1$, therefore $q = 1$ is the *pass-through* point where Renyi entropy and its diversity stop all of their biased favors and weight the local frequencies $c_{ij,l}$ by their *own* values, which is what Shannon entropy captures. This suggests a *meeting point* of the two entropies. Indeed, even though Equations (8), (9) and (10) are *undefined* at $q = 1$, their limits exist when $q \to 1$ (see the proof below) and become the Shannon entropy and the diversity defined in Section 3.2 .

Proof of Renyi Entropy's Limit: At $q = 1$ Equation (8) is undefined at its form $f(q)/g(q) = 0/0$, where $f(q) = -\log \sum_l \left(P_{ij}^l \right)^q$ and $g(q) = 1 - q$. Therefore, we use l'Hôpital's rule to find its limit, which is $\lim_{q\to 1} f(q)/g(q) = \lim_{q\to 1} f'(q)/g'(q)$. $f'(q) = \left(-1/\sum_l \left(P_{ij}^l\right)^q\right) \times \sum_l (P_{ij}^l)^q \log P_{ij}^l$, and $g'(q) = -1$ (assuming natural logarithm for simplicity). Plug in the value $q = 1$ in the equation $f'(q)/g'(q)$, we get:
$\lim_{q\to 1} f(q)/g(q) = f'(q)/g'(q)|_{q=1} = \sum_l P_{ij}^l \log P_{ij}^l$. The last formula is nothing but Shannon entropy, thus the limit existence of Renyi entropy is proved. This also leads to the limit of diversity as $D = exp(H)$, and at $q = 1$ Equations (9) and (10) become Equations (6) and (7), respectively.

We see that the impact of the local frequency on the diversity is not always necessarily determined by just its own value $c_{ij,l}$, but also by the value of parameter q. This moves us one step closer to solving the problem of coincidences, which we are going to discuss in Section 3.4.

3.4 Coincidences

Coincidences occur when two people happen to be at the same places at the same time but never or rarely get a chance to see and communicate with each other, thus less possibility of being friends. This happens often in popular and crowded places where coincidences are frequent, such as cafeteria, public libraries, etc.

Consider the following example: assume there are 5 cells and consider two user pairs (a,b) and (c,d) with co-occurrence vectors: $C_{ab} = (10, 1, 0, 0, 9)$, $C_{cd} = (2, 3, 2, 2, 3)$, respectively. We also assume that cells 1 and 5 are *highly* crowded places, cell 2 is *medium*-crowded, while cell 3 and 4 are *non*-crowded, based on the number of visits. Intuitively, this example suggests that c and d are far more socially connected than a and b as the co-occurrences of a and b are likely coincidences; the co-occurrence at cell 2 would be the only one that is *medium*-significant for friendship of a and b, while c and d would have 7 of such or even more significant co-occurrences from cells 2, 3 and 4. First, obviously, using the total number of co-occurrences would give a wrong suggestion that (a,b) are socially closer to each other than (c,d). Second, using the number of co-occurrence locations (NL) for social strength would give us a relative value $NL_{ab}/NL_{cd} = 3/5$, which still indicates a recognizable level of connection of (a,b) compared to (c,d), but a *fair* measure would reasonably want that level to be low.

Now let's see how diversities of Shannon and Renyi entropies address the challenge in the example above. We set the value of q (order of diversity) to 0.5, less than 1, which, according to the discussion in Section 3.3, will limit the impact of coincidences. The relative value for Shannon entropy of two user pairs is $H_{ab}^S/H_{cd}^S = 0.86/1.59 = 0.54$, relative Shannon's diversity is $D_{ab}^S/D_{cd}^S = 2.35/4.90 = 0.48$, relative Renyi entropy is $H_{ab}^R/H_{cd}^R = 3.20/5.63 = 0.56$, relative Renyi's diversity is $D_{ab}^R/D_{cd}^R = 24.60/279.67 = 0.09$. First, the Renyi's diversity shows a relatively high level of social connection of (c, d)) compared to (a, b) ($D_{ab}^R/D_{cd}^R = 0.09$), which we would expect intuitively. Second, compared to Renyi's diversity, Shannon's diversity does *not* limit the impact of coincidences, consequently, the social strength of (a,b) is still high compared to that of (c,d) ($D_{ab}^S/D_{cd}^S = 0.48$). Third, using entropy (either Shannon or Renyi) instead of diversity as a metric of social strength still results in a relatively high level of connection of (a,b)

compared to (c,d). Therefore, this example confirms our discussion in Section 3.2 that Entropy can only act as the index of diversity, but should not be used as a direct metric for social connection.

We see that coincidences often produce high local frequencies $c_{ij,l}$, which, if misjudged, can be overestimated. However, Renyi entropy and its diversity give us the ability to control the impact of coincidences on diversity through q, which is *sensitive* to the values of local frequencies. q is one of the optimization parameters and will be determined experimentally in Section 5.4.1.

Towards this end, we have focused on eliminating the impact of coincidences at crowded places. However, we have not yet answered the following two questions: **1)** What characterizes crowded and non-crowded places, or even further, the level of crowdedness? **2)** What can be used to determine the likelihood of coincidences in co-occurrences, even when local frequencies are low, and oppositely, the likelihood of non-coincidences when frequencies are low or high? We are going to answer these questions in Section 3.5 utilizing ***Location Entropy*** and ***Weighted Frequency***.

3.5 Location Entropy

Location entropy is a crucial part in **weighted frequency**. It was first introduced in [7] to describe the popularity of a location. Let l be a location, $V_{l,u} = \{< u, l, t >: \forall t\}$ be a set of check-ins at location l of User u and $V_l = \{< u, l, t >: \forall t, \forall u\}$ be a set of all check-ins at location l of all users. The probability that a randomly picked check-in from V_l belongs to User u is $P_{u,l} = |V_{l,u}|/|V_l|$. If we define this event as a random variable, then its uncertainty is given by the Shannon entropy as follow:

$$H_l = -\sum_{u, P_{u,l} \neq 0} P_{u,l} \log P_{u,l} \tag{11}$$

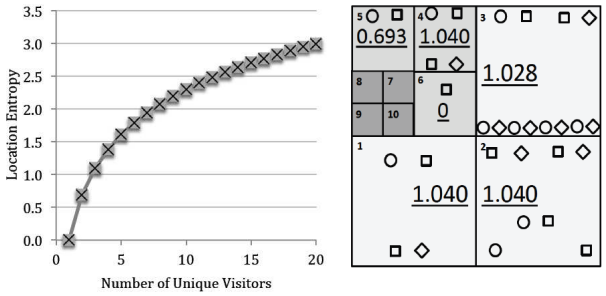
A high value of the location entropy indicates a popular place with many visitors and is *not* specific to anyone. On the other hand, a low value of the location entropy implies a private place with few visitors, such as houses, which are *specific* to a few people.

To help understand the meaning of location entropy, let's assume a simplified case, when N users have visited a location l and each user visited it exactly once. The location entropy then becomes:

$$H_l = -\sum_{u=1}^{N} \frac{1}{N} \log \frac{1}{N} = \log N \tag{12}$$

As we see in this simplified case, location entropy is the logarithm of the number of unique users, who have been at the place. We show the dependence of location entropy's value on the number of unique visitors for this simplified case in Fig 3(a). Fig 3(b) shows an example of location entropy for the case of three users from Fig 1, where the value of location entropy of each cell is underlined. Note that cell 7,8,9 and 10 have no visitors and they have a default value of 0 for location entropy (not shown in the figure). Fig. 3(b) tells us that location entropy is not really determined by the number of visits, but rather by the number of unique visitors. In addition, the location entropy is higher if the location is less specific to any user. Location entropy helps us answer two questions:

- Using location entropy, we can determine the places where coincidences are highly probable, even when the frequency of a user pair in such places is low. That is because location entropy for a place takes into account the visits of all others to that place.
- A low number of co-occurrences (low local frequency) at an uncrowded place can also be a significant indicator of social connections, ***even if*** the diversity is low. When the number of co-locations is low, the diversity will also be low, hence this type of co-occurrences cannot be captured by the diversity measure.



(a) Number of Unique Visitors vs. LE (b) Example of LE

Figure 3: Location Entropy (LE)

Therefore, it is necessary to ensure that co-occurrences at highly private places are given more priority or weight.

Using location entropy, we will introduce weighted frequency in Section 3.6 to capture this type of co-occurrences

3.6 Weighted Frequency

Co-occurrences in small uncrowded places, such as private houses, often results in more social interaction, as compared to those in crowded places. Therefore, the probability of friendships strongly depends on the locations of co-occurrences. Given the co-occurrence vector of Users i and j in Equation (2), we define the ***weighted frequency***, which tells us how important the co-occurrences at non-crowded places are to social strength, as follow:

$$F_{ij} = \sum_{l} c_{ij,l} \times \exp(-H_l) \tag{13}$$

It is interesting to note that $exp(-H_l)$ is the *inverse of diversity* of a location in terms of its visitors. This weighted frequency is inspired by ***tf-idf*** - a numerical statistic widely used in information retrieval and text mining [3] to measure the importance of a term/word t to a document in a corpus. ***tf*** is term frequency and often taken as the number of times the term appears in a document. ***idf*** is inverse document frequency, defined as $|D|/(|d \in D, t \in d| + 1)$ - the total number of documents $|D|$ divided by the number of documents that have t. In our problem, *location* is similar to *document* in ***tf-idf***, thus the number of co-occurrences at a location is similar to *term frequency* in a document. However, to weight co-occurrences, we use $exp(-H_l)$, not ***idf***, since location entropy provides insights into the intrinsic characteristics, i.e., the visiting patterns to a location. ***idf*** is not suitable here because by its definition, it says how important or how specific a user pair is to a location, but we want to answer a different question: how important a co-visit to that location is to a pair of user?

Another tempting approach, which is also inspired by ***tf-idf*** and in fact, used in [7], would be using $c_{ij,l}/\sum_l c_{ij,l}$ - the number of co-occurrences of a user pair at a location divided by the total number of co-occurrences by all user pairs at that location. To show the shortcoming of this approach, assume we have a private living house of a couple, who have made check-ins to produce 1000 co-occurrences. Another guest couple visited them once and made 1 co-occurrence. The $c_{ij,l}/\sum_l c_{ij,l}$ for the guest couple in this house is $1/(1000+1) = 1/1001$, which is very low and, therefore, would say nothing about the social connection of the guest couple, but as we know, such a co-occurrence, even just one, is a high indication of a social connection. Our ***weighted frequency***, however, looks at this case *differently*. Since the house is visited by few people, its location entropy is low (0.0079), which makes a high value of

weight $exp(-H_l) = 0.9921$ (note that $0 < exp(-H_l) \leq 1$). Thus, this only co-occurrence makes a high impact on weighted frequency and is significant for the connection of the guest couple.

To continue the example in Section 3.4, we calculate weighted frequencies for the two couples (a,b) and (c,d). As we assumed earlier, cells 1 and 5 are crowded places. To compute location entropy, we also need the visit information of other users in each cell. To achieve that, assume there are additional 20 visitors at each of cells 1 and 5, and each of them visited the cell 10 times. For simplicity, also assume that each user a, b, c and d visited each cell as many times as they co-occurred with their partners. The weighted frequency for each pair is $F_{ab} = 1.10$ and $F_{cd} = 3.07$. Our analysis shows that F_{ab} is mostly impacted by cell 2, and F_{cd} is mostly impacted by cells 2, 3 and 4, which matches our expectation that only non-crowded places contribute to weighted frequencies.

Note: Diversity and weighted frequency answer two different question. Diversity *decreases* the impact of frequent coincidences while weighted frequency *increases* the impact of co-occurrences at less crowded places; the less crowded, the more impact.

Data sparseness: Weighted frequency plays an important role when it comes to data sparseness, i.e., when the availability of spatiotemporal data is very limited - only few co-occurrences for each couple, the Renyi's diversity can be very low. However, weighted frequency compensates for low diversity by further looking into location characteristics to capture the co-occurrences at non-crowded areas, which can be insignificant for diversity, but very significant for weighted frequency, and for friendship, consequently.

3.7 Social Strength

So far, we have formulated two independent ways, through which co-occurrences contribute to social strength: 1) **Diversity** (through Renyi entropy) - which measures how diverse the co-occurrences of two people are, and at the same time, can control and tell us how much coincidences can impact diversity, and 2) **Weighted frequency** - which favorably captures the local frequencies of co-occurrences at uncrowded places and can compensate for diversity in case of data sparseness. If we want to combine these two measure to produce an ultimate one for social strength, it is necessary to understand the *relative importance* of each component-measure to social strength. To illustrate, from the example discussed in Sections 3.4 and 3.6, we have ($D^R_{ab} = 24.60$, $F_{ab} = 1.10$), ($D^R_{cd} = 279.67$, $F_{cd} = 3.07$). We see that the two measures have different scales; as we decrease the order of diversity q, diversity will scale itself up to more clearly differentiate the impacts of coincidences and non-coincidences. As we see $D^R_{ab}/D^R_{cd} = 0.09$, but at the same time the diversity's scale goes up to 279.67 and weighted frequency remains at a low scale $F_{cd} = 3.07$. This challenge influences us in the way we combine the two measures together.

We now formulate the social strength, which ultimately tells how close two people are, by doing a linear regression over diversity D_{ij} and weighted frequency F_{ij}:

$$s_{ij} = \Phi(D_{ij}) + \Psi(F_{ij}) \tag{14}$$

where Φ and Ψ are two linear functions and s_{ij} is the ultimate strength measure we look for. Since D_{ij} focuses on the distribution of co-occurrences over different locations, while F_{ij} focuses on the intrinsic properties of locations, they are independent of each other, subsequently, Equation (14) takes us to a multiple regression problem over two independent variables D_{ij} and F_{ij}. For convenience of conducting the multiple regression, we rewrite Equation (14) in an explicit form through optimal parameters α, β and γ:

$$s_{ij} = \alpha.D_{ij} + \beta.F_{ij} + \gamma \tag{15}$$

where D_{ij} and F_{ij} are defined in Equations (10) and (13), respectively. Parameters α, β and γ [1] can be learned from dataset and/or provided by user. In Section 5.4.2 we show how these parameters can be learned from training data and applicable across networks.

Equation (15) is the final formula to determine social strength between two users given their co-occurrences in time and space.

4. OPTIMIZATION

To be efficient with massive datasets, we optimize the implementation by using k-d tree [1] for data structure and using MapReduce framework to parallelize the computation.

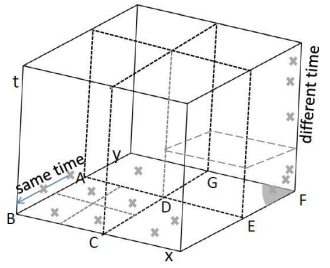


Figure 4: Data Structure

Data structure: We use k-d tree [1] with slight modifications to make it suitable for our problem. It is a 3-D tree shown in Fig. 4, with an (x, y) plane for representing locations, and a t-axis for time. The details of k-d tree's implementation can be found in literature [1]. Our focus here, however, is how to split the tree or sub-tree (named ***rectangular parallelepiped*** or **RP** for short) into smaller sub-trees when inserted data exceeds its capacity? To do that, we can either split the 1-D time interval into two equal halves or the 2-D spatial square of the **RP** into four equal quads. The former method is cheaper as it is a 1-D split. However, consider the case of quad $ABCD$ in Fig 4, spatial splitting is a better option as the spatial points are evenly spread out over the quad but all the temporal points shrink to one end of the time interval. In contrast, in the $DEFG$ quad, temporal splitting is a better option. To facilitate decision making, *one more time*, we use entropy. Assume we have N check-ins in an **RP**. Divide the **RP**'s spatial quad into $S = 4^n$ equal cells indexed from 1 to S, and the RP's time interval into $T = 2^m$ equal sub-intervals indexed from 1 to T (n and m are integers), then for the check-ins, the spatial Shannon entropy in the quad and the temporal Shannon entropy in the time interval are:

$$H_s = -\sum_i \frac{s_i}{N} \ln(\frac{s_i}{N}), \quad H_t = -\sum_j \frac{t_j}{N} \ln(\frac{t_j}{N}) \tag{16}$$

where s_i and t_j are the numbers of check-ins in spatial cell i and time sub-interval j, respectively, and $1 \leq i \leq S$, $1 \leq j \leq T$. The more evenly spread-out the check-ins in the *quad* (or the time interval), the higher the entropy. These two quantities give us the clue of which dimension to split in an **RP** when it reaches capacity N. The empirical values for S and T is 16 ($n = 2$ and $m = 4$). In general, when the two entropies are roughly the same, time splitting is chosen to save storage cost.

Implementation with MapReduce: First, with k-d tree, the search for co-occurrences becomes efficient and standard as the time interval and the spatial quad can efficiently filter out all non-candidate points. The MapReduce implementation can be done in

[1]It is possible to keep only one parameter, say α, let $\beta = 1$ and skip γ. However, we keep all the three parameters just to follow the more traditional form of the multiple regression problem.

two phases. In the first phase, Maps build partial co-occurrence vectors in sub-trees, and Reduce combines them to make full co-occurrence vectors and compute diversities. In the second phase, Maps compute location entropy for each **RP**, while Reduce will use the entropy values to compute weighted frequencies.

5. PERFORMANCE EVALUATION

5.1 Dataset

The data used in the experiments was collected by Gowalla - a location-based social network, where users shared their locations through check-ins. The data was collected from February 2009 to October 2010 and consists of two different sets. The first set is spatiotemporal data, which has 6,442,890 check-ins from 196,591 users. Each check-in has format: <user ID, latitude, longitute, timestamp, location ID>. The second set is a social graph of friendships among users. It has 950,727 edges (or friendships).

Since Gowalla's spatial data are heavily concentrated in the US, we used only the spatiotemporal data within the US for the experiments. We divided the data into two subsets, **training** set and **evaluation** set. The **training** set contains $2,957,830$ check-ins in the West of USA (named L_{west}), and the social network of $102,320$ users who performed check-ins in the West (named S_{west}). The **evaluation** set contains $3,485,060$ check-ins in the East of USA (named L_{east}) and the corresponding social network of $95,725$ users (named S_{east}). Since some users performed check-ins in both the West and East parts, the two subsets S_{west} and S_{east} overlap. However, our analysis shows that the overlaps of users is just 0.74%, and of friendships is only 0.107%. Thus, the overlap is insignificant and cannot affect our evaluation.

5.2 Methodology

The two metrics we use to measure the accuracy of our techniques are precision and recall. Let TC be the set of true social connections reported by Gowalla's social network (i.e., **ground truth**) and RC be the set of user pairs that our model reported as socially connected. The **precision** and **recall** are defined as:

$$P = \frac{|TC \cap RC|}{|RC|}, \quad R = \frac{|TC \cap RC|}{|TC|} \tag{17}$$

where $\cap$ denotes the intersection operation.

5.3 Experiment Setup

We conducted our experiments on Amazon EC2 cluster with instances running on 64 bit Fedora 8 Linux Operating System with 15GB menory, 4 virtual cores and 4 disks with 1,690 GB storage. Our Map/Reduce algorithm is implemented using Hadoop version 0.22.0. To obtain consistent results, we used the same setup to perform all the experiments.

5.4 Results

5.4.1 Order of Diversity

In this set of experiments, we want to examine how the order of diversity q controls the impact of coincidences on diversity and find the optimal value for q. Towards this end, we use *only* diversity as social strength. To be completely unbiased, we do this experiment using only the training data L_{west} and S_{west}.

We perform this particular experiment through the following steps. **Step 1:** Vary the order of diversity q from 0 with a step of 0.1, then for each value of q, we calculate diversity D^q_{ij} based on Equation (10). **Step 2:** Since S_{west} only tells us if two users are friends or not, while our output D_{ij} gives us a numerical value (assumably strength), we need to somehow make the two comparable. To accomplish this, we define the threshold of diversity to be D^q so that: if $D^q_{ij} \geq D^q$ then User i and User j are considered to be friends by our model; otherwise they are not. Therefore, we vary threshold D^q from 0 with a step of $max(D_{ij})/1000$, take user pairs with diversity $D^q_{ij} \geq D^q$ (assuming they are friends) to compare with the real friendship information in S_{west} and calculate precision P and recall R. As a result of varying D^q, we get the dependence of precision and recall on q.

Figs. 5(a) and (b) show the results of how q impacts precision. The x-axis shows the order of diversity q and the y-axis shows the precision. To simplify the graph visualization, we split the graph into two, each shows three curves, each curve corresponds to one level of recall. In addition, we only show the results of q that ranges from 0 to 2.0 to keep a high level of details since further increasing q beyond 2.0 decreases the precision dramatically. We made the following **observations**:

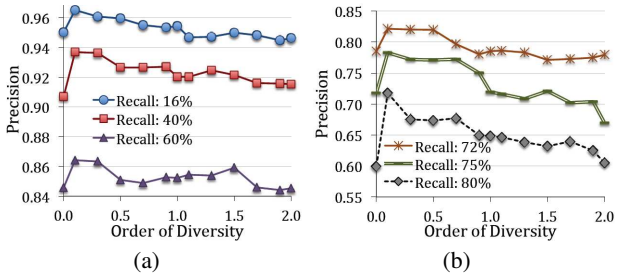


Figure 5: The impact of the order of diversity on precision.

- Our major observation is all the curves at 6 different recall's levels show the same behavior: they all peak at $q = 0.1$, which says $q = 0.1$ is the *optimal value* for limiting coincidences' impact. We believe this optimal value is a general phenomenon across networks for two reasons: first, all the networks nowadays share the same nature of check-ins as users share their locations with their friends; second, coincidences are general spatial phenomena and happen to all networks *without bias* to any particular network. To confirm this phenomenon, we repeated this experiment on another similar dataset from a different network - Brightkite, which consists of 58K users, 214K connections and 4.5M check-ins. The result showed a peak again at $q = 0.1$ with a very insignificant fluctuation (0.004), which can be considered as experimental uncertainty.
- The case $q = 0$ makes the diversity equal to the number of co-occurrence locations (a.k.a *richness*). Fig. 5 shows the fact that simply setting the diversity to the number of co-occurrence locations will produce ***low*** precision. This is because coincidences are completely ignored and all cases of co-occurrences are considered equally important.
- When q increases from 0.1 to 2.0, the diversity ***increasingly favors high*** local frequencies. Consequently it favors coincidences because coincidences often produce high local frequencies. Therefore coincidences now are out of control and have more impact on diversity, which causes false predictions, and consequently, causes the decrease in precision.
- Further decreasing q from 0.1 to 0 also results in the degradation of precision, because low values of q ($q < 0.1$) not only limits the impact of *coincidences* (*high* local frequencies), but also limits the impact of *non-coincidences* (*medium* local frequencies), which results in *excessive* controlling or *over*-limiting.

We will use the optimal value $q = 0.1$ for the rest of our experiments. Note that diversity mainly deals with precision since its role is to avoid coincidences. Therefore in Fig 5 we used precision to learn about the order of diversity. The low recall associated with diversity will be compensated by weighted frequency, which is what we are going to examine next.

5.4.2 Social Strength

Our goals in this set of experiments is 1) to compute the social strength by experimentally conducting multiple linear regression over diversity D_{ij} and weighted frequency F_{ij}; 2) to evaluate the social strength by relating it to friendship information from the ground truth S_{east}.

Linear Regression: In order to find the social strength in the evaluation dataset L_{east}, we first need to use the training set (L_{west} and S_{west}) to learn about the parameters α, β and γ in Equation (15) (See section 3.7). Thus, we need diversity D_{ij}, weighted frequency F_{ij} and strength $\hat{s}_{ij}$ (computed only based on social graph). We already have D_{ij} and F_{ij} computed from Equations (10) and (13). However, S_{west} is a social graph that only tells us if two users are friends or not, but not the strength. Fortunately, there exist different techniques to calculate social strength based *solely* on a social graph. We will use three techniques, which have been shown to have high performance [13], including **Jaccard's index**, **Adamic/Adar similarity** and **Katz score**.

Jaccard's index of Users i and j (J_{ij}) measures the probability that both i and j have a randomly selected person as a friend.

$$J_{ij} = |\Gamma(i) \cap \Gamma(j)|/|\Gamma(i) \cup \Gamma(j)| \tag{18}$$

where $\Gamma(i)$ and $\Gamma(j)$ are the sets of friends of i and j, respectively. Jaccard's index is inversely proportional to the total number of friends of i and j. Thus, the fewer friends that i and j have, the more *influential* their common friends are to Jaccard's index.

Adamic/Adar similarity (AA_{ij}) looks further at the popularity of each common friend of Users i and j and weights each of them differently. Consequently, a common friend, who is also a friend of *many* other people, has less impact on the similarity.

$$AA_{ij} = \sum_{k \in \ \Gamma(i) \cap \Gamma(j)} \frac{1}{\log \Gamma(k)} \tag{19}$$

Katz score K_{ij}: If we represent a social network as a undirected graph with users as nodes and edges as friendships between users, then the Katz score considers the ensemble of all paths between two nodes and sums over this ensemble:

$$K_{ij} = \sum_{l} \varepsilon^l \times |P_{ij,l}| \tag{20}$$

where $P_{ij,l}$ is the set of paths of length l from node i to node j, and ε is a small positive constant that defines the *attenuation* of Katz score as the length l grows. The optimal value of ε is shown in [13] to be of the order 10^{-3}.

Subsequently, $\hat{s}_{ij}$ can be any of the three measures above. Among the three, Katz score has the best performance, followed by Adamic/Adar similarity, and by Jaccard's index [13].

The optimal values of α, β and γ as the results of the *least-square method* in linear regression [2] are given as follow:

$$\alpha = \frac{(\sum F_{ij}^2)(\sum D_{ij}.\hat{s}_{ij}) - (\sum D_{ij}.F_{ij})(\sum F_{ij}.\hat{s}_{ij})}{(\sum D_{ij}^2)(\sum F_{ij}^2) - (\sum D_{ij}.F_{ij})^2} \tag{21}$$

$$\beta = \frac{(\sum D_{ij}^2)(\sum F_{ij}.\hat{s}_{ij}) - (\sum D_{ij}.F_{ij})(\sum D_{ij}.\hat{s}_{ij})}{(\sum D_{ij}^2)(\sum F_{ij}^2) - (\sum D_{ij}.F_{ij})^2} \tag{22}$$

$$\gamma = \overline{\hat{s}_{ij}} - \alpha.\overline{D_{ij}} - \beta.\overline{F_{ij}} \tag{23}$$

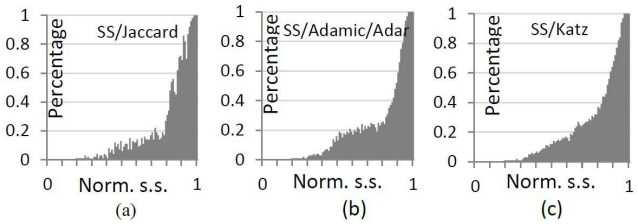


Figure 6: Percentage of real friendships vs. the social strength of buckets.

where $\overline{\hat{s}_{ij}}$, $\overline{D_{ij}}$, and $\overline{F_{ij}}$ are the corresponding mean values of $\hat{s}_{ij}$, D_{ij} and F_{ij}.

Applying each of the three techniques above to the social network S_{west} (i.e., the training data), we compute the social strength $\hat{s}_{ij}$. However, before computing parameters (α, β, γ), we first normalize diversity, weighted frequency and $\hat{s}_{ij}$ so that we can use the values of α and β to analyze the relative importance of each measure (diversity and weighted frequency) to social strength. The values of (α, β) for each of the Jaccard, Adamic/Adar and Katz methods are (0.441, 0.550), (0.476, 0.521), and (0.483, 0.520), respectively. As we see, the two measures are *comparable* in their importance to social strength in all three methods. Weighted frequency gets a slightly higher priority, which reveals a fact that many co-occurrences at uncrowded places have *low* frequencies. These are general phenomena because people check in more frequently at famous and popular places as those are more interesting to share with friends, while uncrowded places are less interesting to share, thus the check-in's frequencies there should be low. It is important and also interesting to note that this nature of check-ins is general to all networks since the main purpose of users' check-ins is to share their locations with friends, despite the fact that different networks might have different ways of encouraging users to perform check-ins. Therefore, consider two scenarios: first, if a partial social network is available explicitly for a spatiotemporal dataset, then it can be used to compute its own parameters (α, β, γ). Second, however, if no explicit social network is available, then the values of (α, β, γ) can be applied across networks without much sacrifice of precision due to the general phenomena discussed above.

Finally, applying these parameters to the evaluation dataset L_{east}, we compute the social strength s_{ij} for new user pairs based on their spatiotemporal data.

Social Strength and Friendship: Our goal now is to show the relationship of our *predicted social strength* with the *friendships*. We do this by *grouping* the user pairs with similar social strength together into subgroups called *buckets* and find the percentage of *real* friendships in each bucket. We perform this experiment through the following steps. **Step 1:** we divide the social strength axis into 100 intervals of equal length $\delta = 0.01$. **Step 2:** we group the user pairs with s_{ij} that belong to the same interval into a bucket. **Step 3:** we take the user pairs in each bucket and check with the social network (i.e., ground truth S_{east}) to find what percents of pairs in each bucket are truly friends as reported by Gowalla.

Fig 6 shows the results in forms of charts. The x-axis shows the middle value of normalized social strength of each bucket (or interval), while the y-axis shows the percentage of real friendships in each bucket as checked with the ground truth S_{east}. The three graphs in Fig 6(a)(b) and (c) correspond to three different cases of which technique is used in the linear regression of social strength in Section 5.4.2: (a) Jaccard's index is used; (b) Adamic/Adar similarity is used; (c) Katz score is used.

Observations: First, as observed in Fig. 6, our predicted social strength is consistent with the ground truth: user pairs with higher social strength have higher percentage of being friends than those user pairs with lower social strength. This also fits the *intuition* that user pairs with high social strength are more involved/interactive with each other, therefore they are more likely to be friends. Second, furthermore, as Katz score is a better metric than the other two [13], our predicted social strength also shows a better performance when Katz score is used in the regression. Fig. 6(c) shows a more consistent curve as the percentage is supposed to increase when the social strength of buckets increases. As Jaccard's index is the worst metric compared to Adamic/Adar similarity and Katz score, we see more fluctuations in 6(a) as the percentage goes up and down, while Fig 6(b) and Fig. 6(c) are smoother, which means more consistent with the ground truth.

5.4.3 Goodness of fit

Our goal in this experiment is to evaluate how well our *predicted* social strength from spatiotemporal data L_{east} fits the *observed* strength computed based only on social graph S_{east}? This is known as *goodness of fit*. This differs from the the previous experiments as in Section 5.4.2 we tested our social strength against friendship, but not observed strength. Hence, we use the **coefficient of determination** R^2 to measure the *variance* of our predicted social strength s_{ij} from $\hat{s}_{ij}$ computed *solely* based on ground truth S_{east} using each of the techniques above.

Coefficient of Determination measures how well the the predicted values fit the observed values. Let N be the number of user pairs, $\overline{\hat{s}_{ij}} = \sum \hat{s}_{ij}/N$ be the mean of the observed social strengths, $SS_{tol} = \sum_i (\hat{s}_{ij} - \overline{\hat{s}_{ij}})^2$ be the total sum of squares and $SS_{err} = \sum_i (\hat{s}_{ij} - s_{ij})^2$ be the sum of squares of residuals. The coefficient of determination is defined as:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tol}} \tag{24}$$

R^2 is a statistic that shows the goodness of fit of the model. It ranges from 0 to 1.0; R^2 near 1.0 indicates that the regression results fit the real data well, while R^2 near 0 indicates the opposite.

Table 2: Coefficient of Determination

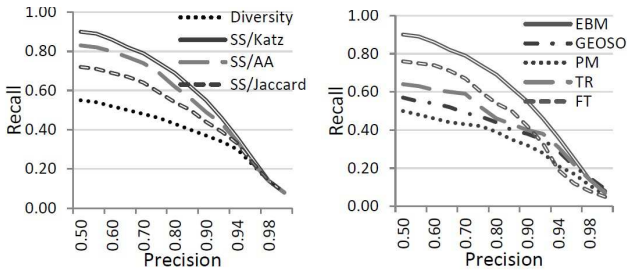
	Jaccard	Adamic/Adar	Katz
R^2	0.691	0.830	0.877

Table 2 shows the values of R^2 for each different technique used to compute $\hat{s}_{ij}$ in the evaluation's social network S_{east}. First, our model very well predicts the social strength. Particularly, if we use Katz score and assume its absolute reliability, then 87.7% of the social strengths in the East of USA are predictable by our model. Second, Katz and Adamic/Adar methods fit our linear regression better than the Jaccard's index. Logically, this also fits the evaluation in [13], which reported that Katz score is a better metric for social strength, followed by Adamic/Adar similarity, and followed by Jaccard's index. Third, the values of R^2 are high, which implies that linear regression is the right choice to integrate diversity and weighted frequency together.

5.4.4 Precision and Recall

We already showed that the precision of EBM is very high in Section 5.4.1, even just using diversity. Here, we would like to evaluate its recall. That is, what percentage of Gowalla's social connections can be predicted by ***just*** analyzing the check-in data? Note that this is a very tough challenge since the check-in data collected from Gowalla is very sparse (unlike more active social-networks such as Foursquare). For example, we analyzed the Gowalla users' co-occurrence vectors and found that out of $996,621$ user pairs who have co-occurred, only 4.3% of them co-occurred more than three times. 95% of pairs have few co-occurrences, which will inevitably limit the opportunity of exploring the social connections from this sparse spatiotemporal data. To alleviate this, we could have used other factors that can help us to infer social connections, such as common friends, common interests, etc. [4]. However, since the focus of this paper is on inferring social connections only from spatiotemporal data, we challenged EBM by limiting its knowledge to only the check-in data. Instead, to slightly level the field, we removed from our dataset those pairs of users who have zero or one co-occurrence and only included the pairs with more than one co-occurrence. This is a fair adjustment as it is almost impossible to infer friendship for those users with one or less co-occurrence (by any method that only relies on the check-in data).

To calculate precision and recall, we need to compare EBM's predictions with ground truth S_{east}. However, EBM gives us numerical strengths, while S_{east} only tells us if two users are friends or not. To work around this, we use the same technique as in Section 5.4.1 by defining a threshold s_0 for social strength and varying it to find precision and recall. Fig. 7(a) shows the results of the EBM's evaluation. The x-axis shows precision and the y-axis shows recall. The dotted line corresponds to the case when only diversity is used as social strength, while the other three are when both diversity and weighted frequency are integrated together to compute social strength, and correspond to three different techniques used in Section 5.4.2.



(a) EBM (b) EBM vs. other models

Figure 7: Precision vs. Recall

Observation 1: It is interesting to note that all the four curves practically meet together when recall is below 35% and start diverging as recall increases. This can be explained as follow: the low recall happens when we set high values for threshold s_0, thus only a *subset* of user pairs with *high* number of co-occurrences can pass through threshold s_0 and be considered. Therefore, handling coincidences is the only essential requirement for such subset. ***Diversity*** satisfies that requirement and it is used in all four cases, thus they all have the same performance level. However, when we reduce threshold s_0 is when user pairs with *fewer* co-occurrences can pass through the threshold and get considered. At this point, just handling coincidences is *not enough* and this is where ***weighted frequency*** comes into play to cope with data sparseness, as discussed in Section 3.6. Weighted frequency takes into account the location characteristics to predict social connections with very *low* number of co-occurrences. Such subtle co-occurrences cannot be captured by diversity, therefore its graph remains below the other three, which do use weighted frequency in addition to diversity.

Observation 2: As shown in Section 5.4.3 and in the related work [13], Katz score is a better metric for ground truth, followed by Adamic/Adar similarity and Jaccard index. Fig. 7(a) shows that

our predicted social strength matches Katz score the best, which enhances the trustworthiness of our model.

Observation 3: Our model achieves both high precision and recall. Particularly, using Katz technique we can achieve (precision, recall) as high as (80%, 70%) or (70%, 82%). Moreover, considering the sparseness of the data where only 4.5% of co-occurred pairs have more than three co-occurrences, being able to get those high precision and recall is a major achievement.

5.4.5 Comparison of EBM with other models

Using the precision vs. recall graphs, we compare the performance of EBM with other four models, which have previously studied the problem, including probability model (PM) [6], GEOSO model [15], the model that utilizes various features (named FT for short) [7], trajectory model (TR) [12]. Fig. 7(b) shows the results.

Observations: *Precision*: EBM outperforms all other four models in precision. This is due to the EBM's ability of controlling the impact of coincidences, which is the real challenge to the other models since coincidences often produce very high local frequencies, which can easily misguide any method that wants to infer social connections from spatiotemporal data. The PM's precision remains the lowest; it assumes that each user can have *at most one* friend, which is almost never true; a possible fix would be removing that assumption, but that would severely interfere with its design and prevent it from formulating the probability of social connection. All PM, GEOSO and TR completely ignore location characteristics (aka *location entropy* in our work), thus there is no easy way for them to handle data sparseness, where *even* coincidences can have low frequencies, which is shown in our work as a difficult and challenging question. TR also does not clearly address coincidences at *high* frequencies. Finally, FT considers coincidences, however, it does not compute social strength but only answers if two people are friends or not. In addition, its ***tf-idf*** fails to capture co-occurrences at private places as shown in Section 3.6.

Recall: Since we challenged all five models with a *highly sparse* dataset, the result of recall truly shows how capable a model is of mining friendship's information. EBM achieves significantly higher recall compared to other models due to its knowledge of the locations (public or private, level of crowded-ness) and applies that knowledge to make *small* things (few co-occurrences) become *significant*. As discussed earlier, PM, GEOSO and TR have no knowledge about locations, thus cannot capture minor co-occurrences, consequently, have low recall. FT considers location characteristics to capture few co-occurrences in uncrowded locations, which explains its higher recall compared to GEOSO, PM and TR, but still lower than EBM's recall. The latter is because with EBM, *diversity* and *weighted frequency* can compensate for each other to avoid coincidences, and at the same time, handle sparse data.

Efficiency: We also parallelized the algorithms for GEOSO in [15], probability model (PM) in [6] and trajectory model in [12] to make the comparison. The feature model (FT) in [7] does not propose any data structure since the model was to target a relatively small proprietary data set, thus we do not examine its efficiency. Using the same setup, we run the experiments using different numbers of nodes (maps). Table 3 shows the number of seconds each model took to run in different setup: 100 nodes, 500 nodes and 1000 nodes. EBM outperforms the other three models. GEOSO and PM use a uniform grid to partition space into equal cells, thus results in high cost of storage as the number of cells grows enormously. TR has high cost in time due to its construction of trajectory (sequence) of user locations. Furthermore, EBM really takes advantages of parallelization as we see the rate, at which the running time decreases when we add more nodes to the computation.

Table 3: Efficiency of EBM and other models

	EBM	GEOSO	PM	TR
100 maps	159.12s	282.36s	203.25s	394.73s
500 maps	34.84s	92.46s	76.92s	129.32s
1000 maps	19.21s	39.22s	29.36s	64.291s

6. RELATED WORK

Given the experimental results are known now, we would like to elaborate on the related studies. Table 4 summarizes EBM and the previous studies with the same problem focus. The check-mark (✓) indicates that the question is addressed by the model, while the cross-mark (✗) indicates the opposite.

Table 4: Questions addressed by EBM and other models

	EBM	GEOSO	PM	FT	TR	BS
Coincidences	✓	✓	✗	✓	✗	✗
Location Characteristics	✓	✗	✗	✓	✗	✗
Data Sparseness	✓	✗	✗	✗	✗	✗
Social Strength	✓	✓	✓	✗	✓	✗

To examine the relationship between a friendship network and the human interactions, Eagle at al. [9] conducted their analysis on two different sets of data of the same group of users: one from mobile phone called "behavioral", another was reported by users called "self report" (shown as **BS** in Table 4). They examined the communications, locations and proximity of the users over an extended period of time, conducted a regression analysis over the data and finally compared the *behavioral* social network to *self-reported* relationships. Their results showed that the two are indeed related. In addition, communication was the most significant predictor of friendships, followed by number of common locations and proximity. However, this early study did not consider the impact of coincidences nor the importance of co-occurrence locations, which has been shown to be significant in our work.

Crandall et al. [6] used a probability model (**PM** in Table 4) to infer the probability of friendships given the co-occurrences in time and space and did the evaluation with a large dataset from Flickr. The first limit of this model is that it makes a simplifying assumption about the structure of the social network: each user can have only one friend, which is not the case in reality. Second, it does not consider the frequency of co-occurrences at each location; all the co-occurrences at one location only count as once. Finally, the impact of coincidences was not addressed, as well as the location characterics, known as location entropy in our work and in [7].

Cranshaw et al. [7] introduced various features such as specificity, location entropy, etc, in order to analyze the social connections (**FT** in Table 4). Their experiments showed that there exists a relationship between the mobility of patterns of a user and the number of the user's friends in the underlying social network. This is an in-depth study and provides much insight into the social network structure. However, they did not consider a subtle question of the social network: how close two people are (*aka* the social strength in our work)? In addition, they studied the location characteristics to avoid coincidences, but for each user pair, the local frequency was not clearly differentiated from location to location. Here, we show that the influences of local frequencies to social connections greatly vary from location to location. Furthermore, our weighted frequency also differs from TFIDF in their work in that we use the *inverse diversity* of a location ($exp(-H_l)$) to weight the local frequency, which considers the visiting patterns to a location and detects the significance of each single co-occurrence to social

strength. Their TFIDF is more to show the specificity of a location to a user pair.

With a similar problem focus, Li et al. [12] also used the history of user locations to develop a similarity measure among users (**TR** in Table 4). They first represented each user as a trajectory in a hierarchical fashion, then used the similarity between the trajectories of two users as their social similarity. The model considers the movements of users in both micro and macro scales. This research is particularly promising for its scalability and its consideration of different level of movements. However, coincidences and location characteristics were not considered, which has been shown to be crucial in our work.

We previously proposed a geometrical model called GEOSO to infer the social connections based on co-occurrences in space and time (**GEOSO** in Table 4). We first defined the social distance geometrically and introduced two properties: *commitment* and *compatibility*, which must be considered by any distance measure. This approach is particularly interesting as the presence of the two properties help avoid coincidences. However, this approach suffers from complexity when working with massive data. In addition, all locations are considered equally important, therefore this is not an ideal approach to apply when it comes to data sparseness when the locations can be significant in predicting social connections.

In this work, we have addressed all the interesting, and at the same time, difficult challenges that the previous studies either ignored or suffered from. Our work does not make or simplify any assumptions and the experiment has proved the high accuracy of our model in predicting social connections with real-world data even when the data is sparse, as well as its efficiency when it comes to the problem of large-scale online processing.

7. CONCLUSIONS

In this paper, we studied the problem of inferring social connections from spatiotemporal data. Towards this end, we presented the EBM model to address some of the subtle questions about social connections, including how to infer the social strength of two people and how to avoid coincidences, which is a challenging problem due to its frequent nature. EBM also alleviated the problem of data sparseness by incorporating the location characteristics into the model when estimating the strength of social connections. Finally, our algorithm is efficient and parallelizable with Map-Reduce framework. Our experiments confirmed the high accuracy and efficiency of the EBM model and its superiority over competitors.

This work opens up new opportunities to answer some of the questions including: How do the social networks influence human physical behavior? How to use the social strengths inferred from spatiotemporal data to further study other aspects of a social network such as its durability and vulnerability? The issues of privacy are also likely to be raised such as how much of spatiotemporal data of a person is enough to maintain the social privacy of that person. These are some of the issues we would like to investigate as part of our future work.

7.1 Acknowledgments

This research has been funded in part by Award No. 2011-IJCX-K054 from National Institute of Justice, Office of Justice Programs, U.S. Department of Justice, as well as NSF grant IIS-1115153, the USC Integrated Media Systems Center (IMSC), and unrestricted cash gifts from Google, Northrop Grumman, and Microsoft. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the sponsors such as the National Science Foundation or the Department of Justice.

8. REFERENCES

[1] J. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[2] C. Bishop. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[4] W. Bukowski, A. Newcomb, and W. Hartup. *The company they keep: Friendships in childhood and adolescence*. Cambridge University Press, 1998.

[5] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD*, KDD '11, pages 1082–1090, New York, NY, USA, 2011.

[6] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.

[7] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, Ubicomp '10, pages 119–128, New York, NY, USA, 2010. ACM.

[8] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[9] N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the NAS*, 106(36):15274–15278, 2009.

[10] M. O. Hill. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54:427–432, 1973.

[11] L. Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

[12] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL*, GIS '08, pages 34:1–34:10, New York, NY, USA, 2008. ACM.

[13] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for IST*, 58(7):1019–1031, 2007.

[14] D. Papadias, Y. Tao, P. Kanis, and J. Zhang. Indexing spatio-temporal data warehouses. In *Proceedings. 18th International Conference on Data Engineering, 2002.*, pages 166–175. IEEE, 2002.

[15] H. Pham, L. Hu, and C. Shahabi. Towards integrating real-world spatiotemporal data with social networks. In *Proceedings of the 19th ACM SIGSPATIAL*, GIS '11, pages 453–457, New York, NY, USA, 2011. ACM.

[16] A. Renyi. On Measures of Entropy and Information. In *Berkeley Symposium Mathematics, Statistics, and Probability*, pages 547–561, 1960.

[17] H. Samet. The quadtree and related hierarchical data structures. *ACM Comput. Surv.*, 16(2):187–260, June 1984.

[18] D. V. Schroeder and H. Gould. An introduction to thermal physics. *Physics Today*, 53(8):44–45, 2000.

[19] H. Tuomisto. A consistent terminology for quantifying species diversity? yes, it does exist. *Oecologia*, 164:853–860, 2010. 10.1007/s00442-010-1812-0.

[20] H. Tuomisto. A diversity of beta diversities: straightening up a concept. *Ecography*, 33(1):2–22, 2010.