# G

## GeoSocial Data Analytics

Cyrus Shahabi and Huy Van Pham
Information Laboratory (InfoLab), Computer Science Department, University of Southern California, Los Angeles, CA, USA

### Synonyms

Friendships; Implicit social connections; Social strength

### Definition

The ubiquity of mobile devices has enabled Location-Based Social Networks (LBSN), such as Foursquare and Twitter, to collect large datasets of people's locations, which tell *who* has been *where* and *when*. Such a collection of people's locations over time (aka ***spatiotemporal*** data) is a rich source of information for studying various social behaviors. One particular behavior that has gained considerable attention in research and has numerous online applications is whether social relationships among people can be inferred from spatiotemporal data and how to estimate the strength of each relationship quantitatively (aka ***social strength***). 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. Thus, the goal is to derive the *implicit* social network of people and the social strength from their real-world location data as opposed to or in addition to their online activities.

***Social strength*** is a quantitative measure between 0 and 1, which shows the extent two people are socially related. The 0 value indicates the absence of a relationship, while 1 indicates the closest possible relationship between two individuals.

**Problem Definition:** Given a set of users $U = (u_1, u_2, \ldots, u_M)$, a set of locations $L = (l_1, l_2 \ldots, l_N)$ 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.

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 (aka ***coincidences***). 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,


S. Shekhar et al. (eds.), *Encyclopedia of GIS*,
DOI 10.1007/978-3-319-23519-6_1566-1

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, the output of the problem is not just a binary value, which indicates whether two people are friends or not, but it must also tell how strong a social relationship is (*social strength*). Finally, there may be a lot of missing data, as people's location data may be sparse.

This entry surveys previous related studies on the considered problem, shows different possible entropy-based solutions, and discusses their pros and cons.

**Terminology:** Terms that are seen frequently in this entry include co-occurrences, coincidences, and frequency. A ***co-occurrence*** indicates that two people are at the same place and the same time. A ***coincidence*** is a co-occurrence between two or more *non-related* individuals by chance, such as people happen to be in a shopping mall at the same time. ***Frequency*** is the total number of co-occurrences of two individuals across all locations.

## Historical Background

There are a number of studies that focused on deriving social relationships from spatiotemporal data. A summary of these studies is provided below.

Eagle et al. (2009) found the relationship between a friendship network and the human interactions by analyzing two different sets of data of the same group of users: one from mobile phone called *behavioral*, and another was reported by users called *self-report*. They examined the communications, locations, and proximity of the users over an extended period of time and compared the *behavioral* social network to *self-reported* relationships. The results showed that the two networks, *behavioral* and *self-reported*, are indeed related. In addition, communication was the most significant predictor of friendships, followed by the number of common relations and proximity.

Crandall et al. (2010) created a model to infer the probability of friendships given the co-occurrences among people in time and space, which is when people were at the same place at the same time. The model was evaluated using a large dataset from Flickr. The first limitation 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 are only counted one. Finally, the impact of coincidences was not considered.

Cranshaw et al. (2010) conducted a regression over various features such as specificity, location entropy, etc., in order to analyze the social connections. 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 in-depth study provides considerable insight into the correlation between geo-spatial and social networks. However, the main limitation of this model is at its output, which is a binary value that only indicates the existence of a friendship, but not a quantitative social strength.

With a similar problem focus, Li et al. (2008) also used the history of user locations to develop a similarity measure among users. They first represented each user as a trajectory in a hierarchical fashion and 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 macroscales. This research is particularly promising for its scalability and its consideration of different level of movements. However, the effects of coincidences and the importance of each location on social strength (aka similarity in this model) were not considered.

Pham et al. (2011) showed that straightforward methods, such as cosine similarity and cross correlation, cannot correctly infer social strength from spatiotemporal data and proposed a geometrical model called GEOSO to infer the

social connections from spatiotemporal data. The model 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, similar to Li et al. (2008), the importance of each location on social strength is not considered. In other words, all locations are considered equally important; therefore, this is not an ideal approach to apply, especially when it comes to data sparseness, which is when the amount of location data per person is small and it is necessary to improve the quality of predicting social strength using different (less obvious) aspects of spatiotemporal data, including the importance of each location.

On the other hand, several studies have focused on the effects of social connection on human movement in real world. Specifically, Cho et al. (2011) observed that up to 30 % of human movement can be shaped by social connections; friends are more likely to perform check-ins at nearby places and the physical distance between two friends follows a power law. Moreover, in two similar studies Scellato et al. (2010, 2011) showed that the average distance between a person and her friends increases with her number of friends, and the chance for two people present together at the same place to be friends is inversely proportional to the total number of people in that place. As you will see later in section "Weighted Frequency", this latter observation is captured by location entropy.

## Scientific Fundamentals

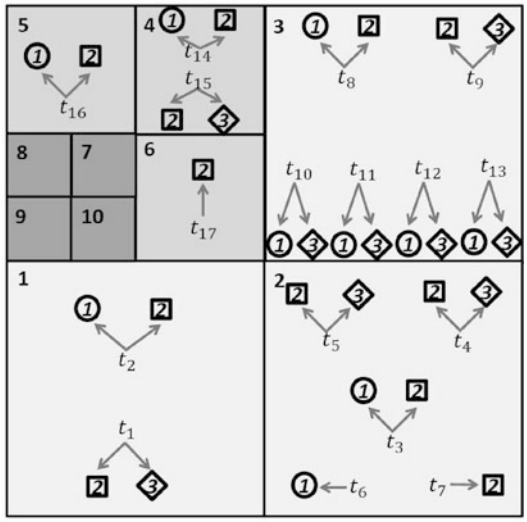
A naive approach to estimate the social strength is to simply count the number of unique locations at which two people co-occurred as their social strength, called *richness*. However, this measure would consider different locations equally important as it ignores the number of co-occurrences at each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

To remedy for these shortcomings, the first possible solution is to use *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. To better handle coincidences, another solution with a more generic entropy, called *Renyi entropy*, is proposed. Renyi entropy is capable of looking at the global pattern of co-occurrences per user pair and has the flexibility of giving more or less weight to coincidences. To further enhance the capability of predicting social strength for those users who have only limited amount of location data (aka *data sparseness*), *weighted frequency* is proposed, which weighs each co-occurrence differently depending on the characteristics (crowdedness) of each individual location. Weighted frequency overlaps with neither Shannon entropy nor Renyi entropy; thus, it can be incorporated with either of them to create a more complete and effective method, which is referred to as EBM (entropy-based model) due to multiple instances of utilizing entropy in the method.

## Preliminary

### Location Representation

We discuss two alternative ways of representing locations. One can use a grid to uniformly partition the space into disjoint cells of *equal size* (Crandall et al. 2010; Pham et al. 2011), 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 co-occurred. Another more effective way is to use a quadtree (Samet 1984). Figure 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




**GeoSocial Data Analytics, Fig. 1** A quadtree storing areas of different levels of popularity, and the visits of users 1, 2 and 3

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 $t_2$.

### 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 $V_1 = (< t_2 >, < t_3, t_6 >, \ldots)$, 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:

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

where $M$ is the number of leaves in the quadtree.

### Co-occurrence Vector

If two users checked in at the same location within a time-interval $\tau$, then we say that they have a co-occurrence. $\tau$ 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:

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

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***.

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

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

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

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:

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

## Models

### 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 (Hill 1973; Jost 2006; Tuomisto 2010a). 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 microstates, in which a system can be (Schroeder and Gould 2000).

As a running example, consider the co-occurrence vectors for three different pairs of users:

*Example 1*

$$C_{12} = (10, 1, 0, 0, 9)$$
$$C_{23} = (2, 3, 2, 2, 3)$$
$$C_{13} = (10, 0, 0, 0, 10)$$

User 1 and user 2 have 20 co-occurrences, and user 2 and user 3 have only 12. However, in the latter case the co-occurrences are spread over five different locations, while in the former case the co-occurrences happened in just three different locations. Similarly, user 1 and user 3 co-occurred only at two different locations. Hence, $C_{23}$ is *more diverse* than $C_{12}$, and $C_{12}$ is *more diverse* than $C_{13}$.

Intuitively, people, who are socially connected, tend to visit *various* places together (Cho et al. 2011; Crandall et al. 2010; Cranshaw et al. 2010; Eagle et al. 2009; Pham et al. 2011). This intuition is captured as *how diverse* their co-occurrences are. Based on the general definition of diversity in Tuomisto (2010b), a formal definition of diversity is given as follows:

**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.

### Social Strength via Shannon Entropy

The method in this section represents social strength via Shannon entropy. First, let's define some notations.

$r_{i,j}^{l,t} =< i, j, l, t >$ 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 follows (the upper index $S$ denotes *Shannon*):

$$H_{ij}^{S} = -\sum_{l} P_{ij}^{l} \log P_{ij}^{l} = -\sum_{l, c_{ij,l} \neq 0} \frac{c_{ij,l}}{f_{ij}} \log \frac{c_{ij,l}}{f_{ij}} \tag{4}$$

where $f_{ij} = \sum_l c_{ij,l}$ is the total number of co-occurrences of user $i$ and user $j$, termed *frequency*, and $P_{ij}^{l} = \frac{c_{ij,l}}{f_{ij}}$ is expressed using the notation of the co-occurrence vector of user $i$ and user $j$. 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* across all locations.

Basically, Shannon entropy shows how diverse a co-occurrence vector is in terms of locations. It is the *unpredictability* of the location of a co-occurrence. That is, it is the amount of location information in the co-occurrences of two users; the more locations, the more information it contains, therefore the less predictable one can make a guess about the location of a co-occurrence, which is randomly picked from the set of all co-occurrences of two users.

To illustrate, Table 1 shows the values of richness $r$, frequency $f$, and Shannon entropy $H^S$ for each co-occurrence vector of Example 1:

**GeoSocial Data Analytics, Table 1** Different candidate measures of social strength for Example 1

| Co-occurrence vector | $r$ | $f$ | $H^S$ | $H^R$ | $D^S$ | $D^R$ |
|---|---|---|---|---|---|---|
| $C_{12} = (10, 1, 0, 0, 9)$ | 3 | 20 | 0.86 | 3.20 | 2.36 | 24.53 |
| $C_{23} = (2, 3, 2, 2, 3)$ | 5 | 12 | 1.59 | 5.63 | 4.90 | 278.66 |
| $C_{13} = (10, 0, 0, 0, 10)$ | 2 | 20 | 0.69 | 0.69 | 1.99 | 1.99 |

**Advantages:** The advantage of Shannon entropy is that its capture of diversity is consistent with the intuitions of friendships. Observe that $C_{23}$ has the highest value of Shannon entropy because it is the most diverse vector, followed by $C_{12}$ and $C_{13}$. From this example and Eq. (4), observe that: **(i)** The more locations, the higher the Shannon entropy. This is intuitive as the more places two users visited together, the stronger their connection. **(ii)** The more uniform the components of the co-occurrence vector are, the higher the entropy. In other words, the more uniform the distribution of the co-occurrences across locations, the higher entropy. This is also intuitive for social strength, because close friends tend to hang out at various places together, thus their co-occurrences should be spread out over many locations, which results in more uniform co-occurrence vectors. Shannon entropy reaches its maximum when all the local frequencies are equal to each other.

**Disadvantages:** Shannon entropy may give higher importance to large components of the co-occurrence vector because each component is weighted by its proportional abundance (Jost 2006). From now on, the term ***outliers*** is used to refer to the large components in the co-occurrence vector that *stand out* from the other components. Specifically, in co-occurrence vector $C_{12} = (10, 1, 0, 0, 9)$, 10 co-occurrences in the first cell is an outlier, which contributes more to the value of Shannon entropy as compared to the single co-occurrence in the second cell. The question is what if the first cell is a crowded area, such as a library, where two students come and study every day; thus, they co-occur frequently, but they may not even know or talk to each other. As mentioned earlier, this is ***coincidence***.

Clearly, Shannon entropy well captures the diversity of co-occurrences, but it cannot limit the impact of coincidences. Therefore, if one is interested in eliminating coincidences, especially when data covers crowded places, a more flexible measure is needed to provide the capability of handling coincidences. The solution to this is Renyi entropy, which is discussed next.

## Social Strength via Renyi Entropy

### Formulation

The purpose of generalizing the method to Renyi entropy is to account for coincidences. The impact of coincidences need to be limited as they often produce high local frequencies $c_{ij}^{l}$ – outliers. The use of Renyi entropy as social strength will give the utility to control how much coincidences can contribute to social strength. In fact, Shannon entropy is just a special case of Renyi entropy.

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

$$H_{ij}^{R} = \left( -\log \sum_{l} \left( P_{ij}^{l} \right)^{q} \right) / (q-1) \tag{5}$$

$$= \left( -\log \sum_{l} \left( \frac{c_{ij,l}}{f_{ij}} \right)^{q} \right) / (q-1) \tag{6}$$

where $q \geq 0$ is the order of diversity. Equation (6) expresses Renyi entropy using the notation of the co-occurrence vector of user $i$ and user $j$, which we have previously used and explained in Eq. (4).

The elegance of using the Renyi entropy comes from the parameter $q$, called the ***order of diversity***, which indicates its *sensitivity* to the local frequency $c_{ij,l}$ (Renyi 1960). Specifically:

- **(i)** When $q > 1$ the Renyi entropy $H_{ij}^{R}$ considers the *high* values of $c_{ij,l}$ more favorably. In other words, the higher the local frequency $c_{ij,l}$, the more contribution to the Renyi entropy or the more impact the outliers can make on the social strength.
- **(ii)** When $q < 1$, instead, the Renyi entropy gives more weight to the *low* local frequencies $c_{ij,l}$.
- **(iii)** When $q = 0$, the Renyi entropy is completely ***insensitive*** to $c_{ij,l}$ and gives the pure number of co-occurrence locations – a.k.a. *richness*.
- **(iv)** Case $q = 1$: As we know by now, the Renyi entropy favors local frequencies $c_{ij,l}$ in opposite directions when $q < 1$ versus when $q > 1$; therefore, $q = 1$ is the

*crossover* point where Renyi entropy stops all of its bias and weighs the local frequencies $c_{ij,l}$ by their *own* values, which is what Shannon entropy captures. That is, at $q = 1$ the Renyi entropy becomes the Shannon entropy. Indeed, even though Eqs. (5) and (6) are *undefined* at $q = 1$, their limits exist when $q \to 1$ and become the Shannon entropy defined in section "Social Strength via Shannon Entropy."

To continue the running example (Example 1), Table 1 shows the values of Renyi entropy ($H^R$) with $q$ set to 0.5.

**Advantages:** The advantage of Renyi entropy is its flexibility to limit or increase a particular behavior in co-occurrences. Particularly, it can reduce the impact of coincidences by setting parameter $q$ to low values. An optimal value of $q$ can be obtained experimentally if a ground truth is available. Refer to Pham et al. (2013) for how to obtain the optimal order of diversity when an explicit social network is available.

The impact of the local frequency on Renyi entropy not only depends on the value $c_{ij,l}$ itself but also depends on parameter $q$. This means we are now one step closer to being able to discuss the solution to the problem of coincidences using the order of diversity – parameter $q$. However, in order to aid the discussion of coincidences, we first need to explain the actual meaning of diversity.

### Diversity as Social Strength

We briefly discussed the concept of diversity in section "Diversity in Co-occurrences." However, we have not yet formulated it. In this section, we formulate diversity, discuss its actual meaning and its relationship to social strength.

Even though entropy (either Shannon or Renyi) shows how diverse the co-occurrences are in terms of location, there exists a distinction between *diversity* and *entropy*. Specifically, entropy is often regarded as the *index* of diversity, but not diversity itself. Jost (2006) compared the *role* of entropy as the radius of a sphere, while the role of diversity as the volume of a sphere, where radius is an index used to calculate volume. They showed a general relationship between entropy $H$ and diversity $D$ as follows:

$$D = \exp(H) \tag{7}$$

Therefore, entropy is a reasonable index of diversity, but one should not consider that entropy is diversity.

To illustrate, consider an example of two pairs of users $(a, b)$ and $(c, d)$. Assume their co-occurrence vectors are $C_{ab}$=(1, 1, 1, 1, 1, 1, 1, 1) and $C_{cd}$=(1, 1, 1, 1, 0, 0, 0, 0). It is reasonable and natural to say that $C_{ab}$ is twice as diverse as $C_{cd}$, or in other words, the $C_{ab}$'s diversity is two times $C_{cd}$'s diversity. To clarify the difference between entropy $H$ and diversity $D$, let's compute them for these two co-occurrence vectors. For simplicity, we use Shannon entropy for illustration. Correspondingly, $H_{ab} = 2.079$, $H_{cd} = 1.386$, while $D_{ab} = 8$, $D_{cd} = 4$. Clearly, it is diversity $D = \exp(H)$ that correctly characterizes the meaning of how diverse each co-occurrence vector is: $D_{ab} = 2D_{cd}$, while entropy does not possess the same property: $H_{ab} \neq 2H_{cd}$. For this reason, diversity (exponential of entropy) is often referred to as the ***effective number of states*** (in Statistical Mechanics (Schroeder and Gould 2000)) or ***effective number of species*** (in Ecology (Jost 2006)).

In sections "Social Strength via Shannon Entropy" and "Social Strength via Renyi Entropy" we temporarily used the direct value of Shannon or Renyi entropy as social strength because we needed the concept of entropy before we can introduce diversity formally. Hereafter, as the actual meanings of entropy and diversity have been clarified, and since we aim to characterize social strength by the diversity of co-occurrences in terms of locations, we will use ***diversity*** $D = \exp(H)$, instead of entropy, as *social strength*. Subsequently, we can express diversity in Eq. (7) via the notation of the co-occurrence vector. Combining Eqs. (6) and (7), we have:

$$
\begin{aligned}
D_{ij} &= \exp(H_{ij}^R) \\
&= \exp\left[\left(-\log\sum_l \left(P_{ij}^l\right)^q\right)/(q-1)\right] \\
&= \left[\exp\left(\log\sum_l \left(P_{ij}^l\right)^q\right)\right]^{1/(1-q)} \\
&= \left[\sum_l \left(P_{ij}^l\right)^q\right]^{1/(1-q)} \qquad (8) \\
&= \left[\sum_{l,c_{ij,l}\neq 0} \left(\frac{c_{ij,l}}{f_{ij}}\right)^q\right]^{1/(1-q)} \qquad (9)
\end{aligned}
$$

The upper index $R$ denotes *Renyi*. Diversity for Shannon entropy can be formulated in a similar way. To complete the running example, Table 1 shows the values of richness $r$, frequency $f$, Shannon entropy $H^S$, Renyi entropy $H^R$, Shannon diversity $D^S = \exp(H^S)$, and Renyi diversity $D^R = \exp(H^R)$, all in one combined table for the three co-occurrence vectors $C_{12}$, $C_{23}$, and $C_{13}$.

Observe that Renyi diversity brings up the values to a larger scale to clearly differentiate the social strength of each user pair. Users 2 and 3 co-occurred more uniformly throughout all locations and their local frequencies stay moderate – none of them look like outliers, and consequently they are less prone to coincidences; therefore, their Renyi diversity's value is significantly higher than the values of the other two couples.

Now we are ready to discuss how Renyi entropy and its corresponding diversity treat coincidences.

### 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, shopping centers, etc.

Consider the following example: assume there are five cells and consider two user pairs ($u_1$,$u_2$) and ($u_2$,$u_3$) with co-occurrence vectors: $C_{12} = (10, 1, 0, 0, 9)$ and $C_{23} = (2, 3, 2, 2, 3)$, 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 $u_2$ and $u_3$ are far more socially connected than $u_1$ and $u_2$ as the co-occurrences of $u_1$ and $u_2$ are likely coincidences; the co-occurrence at cell 2 would be the only one that is *medium* significant for friendship of $u_1$ and $u_2$, while $u_2$ and $u_3$ 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 (aka *frequency*) would give a wrong impression that ($u_1$,$u_2$) are socially closer to each other than ($u_2$,$u_3$). Second, using the number of co-occurrence locations ($R$) (aka *richness*) for social strength would give us a relative value $R_{12}/R_{23} = 3/5$, which still indicates a recognizable level of connection of ($u_1$,$u_2$) compared to ($u_2$,$u_3$), 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 "Social Strength via Renyi Entropy", will limit the impact of coincidences. The relative value for Shannon entropy of two user pairs is $H_{12}^S/H_{23}^S = 0.86/1.59 = 0.54$, relative Shannon's diversity is $D_{12}^S/D_{23}^S = 2.35/4.90 = 0.48$, relative Renyi entropy is $H_{12}^R/H_{23}^R = 3.20/5.63 = 0.56$, and relative Renyi's diversity is $D_{12}^R/D_{23}^R = 24.60/279.67 = 0.09$. **First**, the Renyi's diversity shows a relatively high level of social connection of $(u_2, u_3)$ compared to $(u_1, u_2)$ ($D_{12}^R/D_{23}^R = 0.09$), 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 ($u_1$,$u_2$) is still high compared to that of ($u_2$,$u_3$) ($D_{12}^S/D_{23}^S = 0.48$). **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 ($u_1$,$u_2$) compared to ($u_2$,$u_3$).

Therefore, the example above demonstrates two confirmations that we made earlier: **(i)** Coincidences often produce high local frequencies $c_{ij,l}$, which, if misjudged, can be overestimated. However, Renyi entropy $H^R$ and its corresponding diversity $D^R = \exp(H^R)$ give the ability to control the impact of coincidences on diversity through $q$, which is *sensitive* to the values of local frequencies. **(ii)** $q$ is one of the optimization parameters and can be determined experimentally (Pham et al. 2013).

The distinction between entropy and diversity is clear, and in the end it is diversity that reflects social strength naturally and correctly.

**Disadvantages:** Renyi entropy and consequently its corresponding diversity consider all locations equally important. That is, a co-occurrence at a private house contributes similarly to social strength as a co-occurrence at a crowded area (such as a shopping center) does. This becomes even more significant in inferring social strength when the input data is sparse, i.e., when we have just few co-occurrences per pair of users. A few co-occurrences at crowded places may not tell much about the social strength between the users, but they can be significant to the social strength if they happened at private and/or non-crowded places. To accomplish this thorough consideration of locations, another method called *weighted frequency* is proposed next, which does not overlap with neither Shannon entropy nor Renyi entropy; thus, an incorporation between them is possible.

### Weighted Frequency

This section proposes weighted frequency as social strength. Since weighted frequency uses location entropy mainly in its formulation, a brief discussion of location entropy is provided below.

Location entropy was first introduced in Cranshaw et al. (2010) 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 follows:

$$H_l = -\sum_{u, P_{u,l} \neq 0} P_{u,l} \log P_{u,l} \qquad (10)$$

A high value of the location entropy indicates a popular and crowded place with many visitors, and the number of visits is uniformly spread over the visitors, or in other words, no one made the majority of the visits. 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 few people.

For example, consider two locations $l_1$ and $l_2$ and two users $u_1$ and $u_2$.

*Example 2*

| | $u_1$ | $u_2$ | Location Entropy |
|---|---|---|---|
| $l_1$ | 1000 | 1 | 0.35 |
| $l_2$ | 500 | 500 | 0.69 |

Location $l_1$ was visited by user $u_1$ 1000 times, but by user $u_2$ only once. Location $l_2$ was visited by each user equally, 500 times by each user. Therefore, location $l_1$ is less crowded and more private or more specific to user $u_1$, while location $l_2$ is more crowded and not specific to any user. This observation is indeed characterized by Location entropy $LE$: $LE(l_1) = 0.35$, $LE(l_2) = 0.69$. For a location, the more crowded, the higher location entropy; the less crowded or/and the more private to any user, the lower location entropy.

Note that simply using either (i) the number of unique visitors or (ii) the number of visits to each location would not correctly describe the crowdedness of a location. To illustrate, in Example 2, both locations have two unique visitors, and both have almost the same numbers of visits (1001 and 1000), but location $l_1$ is clearly less crowded as

it is visited only by user $u_1$ most of the time. For more details about location entropy, refer to Cranshaw et al. (2010) and Pham et al. (2013).

***Weighted frequency***, which tells us how important the co-occurrences at non-crowded places are to social strength, is defined as follows:

$$F_{ij} = \sum_{l} c_{ij,l} \times \exp(-H_l) \tag{11}$$

Crowed locations have high location entropy $H_l$, thus low $\exp(-H_l)$, and consequently the impact of $c_{ij,l}$ is decreased. On the other hand, for non-crowded locations, $\exp(-H_l)$ is high and this increases the impact of $c_{ij,l}$.

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 (Blei et al. 2003) 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 this 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 weigh co-occurrences, $\exp(-H_l)$ should be used, 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 need to answer a different question: how important a co-visit to that location is to a pair of user?

**Advantages:** The advantage of using weighted frequency as social strength is it differentiates locations from each other and can increase the impact of each location on social strength depending on how popular or crowded it is. The less crowded, the more impact.

**Disadvantages:** The disadvantage is that it cannot capture the diversity of co-occurrences as Shannon entropy or Renyi entropy does. However, its focus is on the popularity of each location, while Shannon entropy or Renyi entropy focuses on the overall picture of co-occurrences across different locations without looking at each location individually; thus, they are two non-overlapping aspects of co-occurrences and an incorporation between weighted frequency and either type of the entropies is possible. This is discussed in the next section.

### Social Strength

An incorporation between Renyi entropy (or Shannon entropy) with weighted frequency can be made in different ways and can be application dependent. These include linear and nonlinear regression between the two components of the regression. To demonstrate, a linear regression is presented below.

Let $s_{ij}$ be the ultimate social strength that includes both diversity and weighted frequency. A linear regression is expressed as follows:

$$s_{ij} = \Phi(D_{ij}) + \Psi(F_{ij}) \tag{12}$$

where $\Phi$ and $\Psi$ 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, and subsequently, Equation (12) 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 (12) in an explicit form through optimal parameters $\alpha$, $\beta$, and $\gamma$:

$$s_{ij} = \alpha.D_{ij} + \beta.F_{ij} + \gamma \tag{13}$$

where $D_{ij}$ and $F_{ij}$ are defined in Eqs. (10) and (13), respectively. Parameters $\alpha$, $\beta$ and $\gamma$ can be learned from dataset and/or provided by user (It is possible to keep only one parameter, say $\alpha$, let $\beta = 1$, and skip $\gamma$. However, we keep all the three parameters just to follow the more traditional form of the multiple regression problem.). As

a good practice, $s_{ij}$ is generally normalized to [0, 1]. For more information about how to obtain the parameters of the regression, refer to Pham et al. (2013).

## Key Applications

### Background for Other Social Studies

Many social studies rely on a weighted social graph. For example, the social influence maximization needs a weighted social graph to model the information propagation in a network (Kempe et al. 2003). Dynamicity of a network focuses on how a social network changes over time and which relationship becomes stale, which still lasts, and which formed recently. An explicit social network may not be fully available (such as for Foursquare) or may not be applicable for studying the dynamicity because stale relationships are not removed from a social network, such as Facebook. On the other hand, relationships based on recent location activities are highly suitable for this problem because they are based and inferred from the recent activities of users.

### Recommendation Engines and Target Advertising

The inferred social connections can be used as suggestions of new online friendships for the purpose of expanding existing social networks, such as Facebook. They can also be used in location recommendations and target advertising, which deliver ads based on explicit and implicit social connections.

### Criminal Investigation

An inferred social connection from spatiotemporal data suggests that two individuals are socially related in real life. It suggests they have been to the same places together, thus possibly committed crimes together. Thus, new or unknown members of a criminal gang can be identified if they are found to be in the implicit/inferred relationships with the members that have already been identified before.

### Epidemiology: Spread of Disease

Another major application of social strength is in Epidemiology where certain types of disease can spread through human contacts, such as Ebola and Influenza. People who are physically close with patients of these types of diseases have a higher risk of contracting them. A high value of the implicit social strength between a person and a patient indicates that the person is at a high risk because the social strength is based on co-occurrences. Consequently, the implicit social strength derived from spatiotemporal data can be used to establish necessary procedures for people with high risks in epidemiology, such as proper quarantines of 21 days for people with risks of contracting Ebola.

## Future Directions

This work opens up new opportunities to answer some interesting questions including: How do 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 possible future directions to continue this line of research.

## References

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD (KDD'11), New York, pp 1082–1090

Crandall DJ, Backstrom L, Cosley D, Suri S, Huttenlocher D, Kleinberg J (2010) Inferring social ties from geographic coincidences. Proc Natl Acad Sci 107(52):22436–22441

Cranshaw J, Toch E, Hong J, Kittur A, Sadeh N (2010) Bridging the gap between physical location and online social networks. In: Proceedings of the 12th ACM international conference on ubiquitous computing (Ubicomp '10), New York. ACM, pp 119–128

Eagle N, Pentland A (Sandy), Lazer D (2009) Inferring friendship network structure by using mobile phone data. Proc Natl Acad Sci 106(36):15274–15278

Hill MO (1973) Diversity and evenness: a unifying notation and its consequences. Ecology 54:427–432

Jost L (2006) Entropy and diversity. Oikos 113(2):363–375

Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, Washington, DC, pp 137–146

Li Q, Zheng Y, Xie X, Chen Y, Liu W, Ma W-Y (2008) Mining user similarity based on location history. In: Proceedings of the 16th ACM SIGSPATIAL (GIS '08), New York. ACM, pp 34:1–34:10

Pham H, Hu L, Shahabi C (2011) Towards integrating real-world spatiotemporal data with social networks. In: Proceedings of the 19th ACM SIGSPATIAL (GIS '11), New York. ACM, pp 453–457

Pham H, Shahabi C, Liu Y (2013) Ebm: an entropy-based model to infer social strength from spatiotemporal data. In: Proceedings of the 2013 international conference on management of data. ACM, New York, NY, pp 265–276

Renyi A (1960) On measures of entropy and information. In: Berkeley symposium mathematics, statistics, and probability, Berkeley, CA, pp 547–561

Samet H (1984) The quadtree and related hierarchical data structures. ACM Comput Surv 16(2):187–260

Scellato S, Mascolo C, Musolesi M, Latora V (2010) Distance matters: geo-social metrics for online social networks. In: Proceedings of the 3rd conference on online social networks, Boston, MA, pp 8–8

Scellato S, Noulas A, Lambiotte R, Mascolo C (2011) Socio-spatial properties of online location-based social networks. ICWSM 11:329–336

Schroeder DV, Gould H (2000) An introduction to thermal physics. Phys Today 53(8):44–45

Tuomisto H (2010a) A consistent terminology for quantifying species diversity? Yes, it does exist. Oecologia 164:853–860. doi:10.1007/s00442-010-1812-0

Tuomisto H (2010b) A diversity of beta diversities: straightening up a concept. Ecography 33(1):2–22