Inferring Real-World Relationships from Spatiotemporal Data

Cyrus Shahabi   Huy Pham
Integrated Media Systems Center
University of Southern California
Los Angeles, CA 90089-0272

Abstract

The pervasiveness of GPS-enabled mobile devices and the popularity of location-based services have generated, for the first time, massive data that represents the movements of people in the real world at a high resolution, aka spatiotemporal data. Such collections of spatiotemporal data constitute a rich source of information for studying various social behaviors, and in particular, give a boost to the study of inferring the real-world social connections from spatiotemporal data. This article surveys the prominent techniques proposed for deriving social connections and social strength from spatiotemporal data and discusses their formulations.

1 Introduction

Social networks have been studied by social scientists since the pre-Internet era, and their relevance particularly increased in the last decade. We identify three periods in the study of social networks corresponding to the growth in the availability of data over time.

The very first period in social networks started back in 1970s [12] when social scientists realized that it was critical to understand the underlying network that portrays people's social connections and influence relationships. Such information is significant in the analysis of the propagation of information, innovations, practice, rumors and contagious infections, and also in commerce including target advertising and recommendations. However, in the pre-Internet era, the problem of identifying "*who is friend of whom*" was challenging, and studies on social networks in this earlier stage had to confine themselves to extremely small datasets [11], which mostly came from some social surveys at very limited scales.

The second period started along with the Internet revolution in the '90s through the development of web, when our lives have continually expanded to occupy virtual worlds [7]. Towards the end of the last decade, the research on social networks witnessed an explosion. To a large extent, this has been fueled by the spectacular growth of social media and online social networks, such as LinkedIn, Facebook and Twitter, which started in 2003, 2004 and 2006, respectively [11]. These giant networks have produced and continue to produce enormous datasets about hundreds of millions of online connected users in the form of social graphs. Therefore, the "*who is friend of whom*" question, which was a big challenge during the first period, suddenly became a cakewalk. The readily available social graphs collected from online social networks motivated social scientists to move far

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

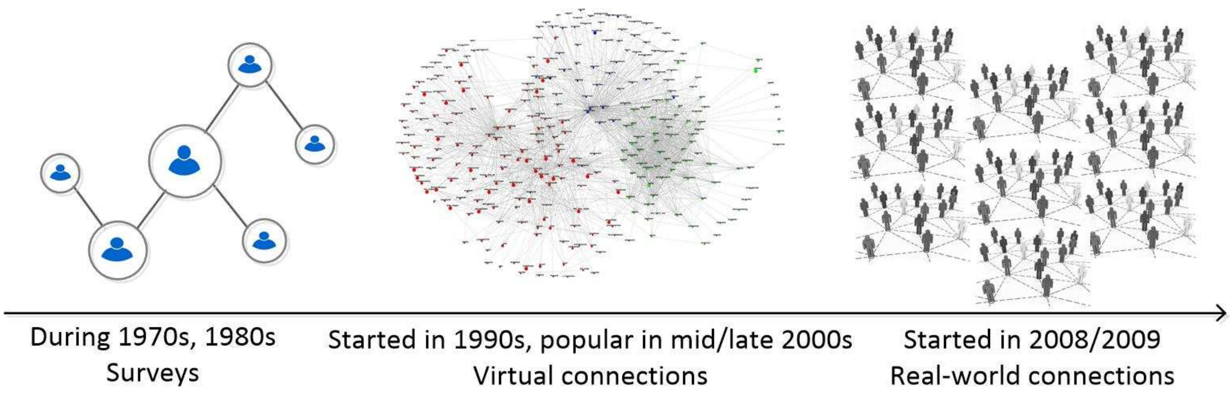


Figure 1: Three periods in social network studies.

beyond the basic question of "*who is friend of whom*" to much more interesting and sophisticated topics. As a result, a large number of studies has been devoted to new questions/solutions related to social networks, including measuring friendships quantitatively [13], identifying most influential people in a network [9], maximizing and speeding up the propagation of information and innovations in a social graph [10], and analyzing the structures and properties of a social network (e.g., density, clusters, stability, etc.) [15] [14]. However, all these achievements may still be considered inadequate in the eyes of the social scientists due to the gap that exists between online social networks (aka the *virtual world*) and the real lives (aka the *real world*). The large volume of studies during this period focused on the virtual world and utilized data collected from online networks. However, the people's relationships in the virtual world may not necessarily correspond to those in the real world.

Subsequently, we are now witnessing the third period as the phase of bridging the gap between the virtual world and the real world. Indeed, the pervasiveness of GPS-enabled mobile devices, and the fact that all the giant social networks have also gone mobile, has introduced massive data that represents the movements of people in the real world at high resolution, specifically by indicating *who* has been *where* and *when* (aka *spatiotemporal data*). Spatiotemporal data can be collected effortlessly from online services, such as geo-tagged contents (tweets from Twitter, pictures from Instagram, Facebook and Flickr, check-ins from Foursquare), or from mobile apps' data (WhatsApp, Glancee), etc. Such collections of spatiotemporal data constitute a rich source of information for studying and inferring various social behaviors, including social connections. For example, for social connections, 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. Since these social connections are inferred from people's real world locations, they constitute social connections that occur in the real world, as opposed to those that may take place only in the virtual world.

The goal of this article is to survey the techniques pursuing the inference of the real-world social connections from spatiotemporal data during, what we called earlier, the third period of social networks.

2 Motivation

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. Just a few years ago, it was practically impossible to find any data that could describe people's locations at high resolution and large scale. However, this is no longer the case nowadays since smart phones and Location-Based Services have produced a tremendous corpus of rich spatiotemporal data. For example, Twitter and Foursquare reportedly received millions of spatiotemporal records per day as geo-tagged tweets or check-ins [16]. This newly available location data is useful for investigating various social behaviors, and thus has motivated social scientists to study and to extend the conventional concept of social behaviors to capture people's activities in the real world, particularly by inferring the implicit networks of social connections based on the actual physical locations of people.

Furthermore, applications for such physically inferred networks of social connections are plenty. First, they include 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). In addition, the physically inferred social connections also have their own unique applications due to the geo-spatial properties. For example, the inferred social connections 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.

3 Challenges

Inferring the implicit social connections is challenging for several reasons.

First, it is not clear what attribute of spatiotemporal data should be measured to infer social connections? If the frequency of co-occurrences (number of times that two people are seen together) is used as the indication of a social connection, one may arrive at a wrong conclusion about their social relationship. To illustrate, suppose two students study at the same library around the same time every day, which results in high frequencies of co-occurrences, but they may not even know each other. This erroneous conclusion can be attributed to coincidences - the fact that the library is a popular location and many students may co-occur frequently there by accident, and thus, the observation that two people only co-occur at the library is not a strong indication of a social connection. On the other hand, a few co-occurrences between two people in a small private place are perhaps a better indication of a friendship. Or alternatively, several co-occurrences at different popular places (e.g., coffeehouses, restaurants) may also be a better indication of friendships.

Second, it is of great interest to quantify social connections, and thus, the goal of inferring social connections from spatiotemporal data is not just to answer the true/false question, whether two people are friends with each other or not? It is more informative to infer a quantitative value that characterizes how strong a social connection is (aka ***social strength***). Lastly, spatiotemporal data is often extremely large, in the order of gigabytes of text, which could render the inference algorithms inefficient, taking too much time and/or resources to perform.

4 Solutions

In this section, we survey the methods proposed for inferring social connections from spatiotemporal data.

4.1 The report-based study

Eagle et al. were among the pioneers to look into the correlation between the location behaviors of users and their social connections. Specifically, in an early study [1], they conducted an analysis on two different sets of data of the same group of users, who were students at a university campus. One dataset collected from mobile phone, called “behavioral”, which contained various features of user data, including the spatial proximity of users at work, their proximity on a specific night of the week, the phone communications between the users and the number of unique locations they were seen together [3] [1]. On the other hand, the other dataset was reported by users themselves, called “self-report”, in which each user indicated who were his/her actual friends. Subsequently, a regression analysis was conducted over the behavioral dataset to find out possible friendships, which in turn were compared with the self-reported friendships. Their results showed that the social relationships extracted from the behavioral dataset were indeed related to the self-reported relationships. In addition, communications were the most significant predictor of friendships, followed by the number of common locations and spatial proximity.

4.2 Probability model

Crandall et al. [4] created a probability model to infer the probability of a friendship between two people given their co-occurrences in time and space. Specifically, they divided the surface of the earth into N grid-like cells, whose side lengths span s degrees of latitude and longitude. Two users are said to co-occur if they were present within the same cell within t days from each other. The number of unique locations (cells) of the co-occurrences between two people is the only factor used to determine the probability of their friendship. Multiple co-occurrences between two people within the same cell are not considered. Hence, the question becomes: What is the probability that two people have a social connection, given that they have co-occurrences in k distinct locations at a temporal range of t?

To formulate the friendship probability, assume that there are M people, each has one social tie, meaning one friend, and the social graph consists of $M/2$ disjoint edges. Each day, each pair of friends chooses to visit a place (i) *together* with probability β, and (ii) *separately* with probability $1-\beta$, with random choices of location. Let F denote the event that they are friends, and let C_k denote the event that they visit k unique locations together on k consecutive days. Consequently, the conditional probability $P(F|C_k)$ indicates the probability of two users being friends given that they co-occurred in k different locations on k consecutive days, which can be expressed by the Bayes' law and has the final formula as follows:

$$P(F|C_k) = \frac{P(F)P(C_k|F)}{P(C_k)} \tag{1}$$

$$= \frac{1}{M}e^{k \log \beta(N-1)+1} \tag{2}$$

The final formula in Equation 2 is obtained after computing the component factors in Equation 1: $P(F)$, $P(C_k|F)$ and $P(C_k)$, the details of which can be found in [4].

The advantage of this model is that it has a final, concise and simple expression for the friendship probability. The model only considers the number of unique locations where two users co-occurred, therefore it reduces the complexity of the algorithm significantly. The authors showed that even very few co-occurrences could lead to a sharp increase in the probability of a friendship, and this finding shows potential implications for the privacy of users on social media sites, which tells how much of the user data can be released until their privacy becomes exposed.

Despite achieving some promising results, the model still has several limitations. The first limitation is the simplifying assumption about the structure of the social network: each user can have only one friend, which is usually not the case in reality. Second, the model does not consider the frequency of co-occurrences at each location; all the co-occurrences at one location count only once. Finally, the issue related to coincidences was not addressed, that is whether the co-occurrences between two people are an indication of a social connection, or are simply coincidences between two people in time and space?

4.3 Trajectory-based model

Li et al. proposed the HGSM model that measures the similarity between two users based on the similarity between their trajectories [2]. The primary idea of this model is that the more similar the location histories of two users are, the more similar their common interests and preferences are, and thus the more likely that they are related socially. The HGSM model (Hierarchical-Graph-Based Similarity Measurement) first represents each user's location history as a trajectory (both sequentially and hierarchically), and the similarity between the trajectories of two users indicates their social similarity.

4.3.1 Trajectory

The *sequential* aspect of the trajectory allows the representation of discrete points in space (the locations visited by a user) as a continuous sequence. On the other hand, the *hierarchical* aspect allows for finer levels of geo-spatial granularity in a trajectory. For example, when a a *stay point* by a user contains multiple businesses located near each other, that stay point can be further turned into a subsequence of points with a smaller scale and a different (finer) level of granularity, and all such levels are said to form a hierarchy of granularity of location history of a user. A stay point is represented as a *cluster* of points at a smaller scale. Therefore, in HGSM, a user's trajectory consists of a set of graphs $HG = \{G\}$ built on different geo-spatial scales of the hierarchy, where each graph $G_i(V, E) \in HG$ is a set of vertexes $V = \{C\}$ (a set of clusters containing the user's stay points) and edges E.

The trajectories (aka sequences) of two users are presented as follows:

$$seq1 =< a_1(k_1) \overset{\Delta t_1}{\rightarrow} a_2(k_2) \overset{\Delta t_2}{\rightarrow} ...a_m(k_m) >$$

$$seq2 =< b_1(k_1') \overset{\Delta t_1'}{\rightarrow} b_2(k_2') \overset{\Delta t_2'}{\rightarrow} ...b_m(k_m') >$$

where $a_i \in V$ is the cluster ID, k_i is the set of time the user successively visited cluster a_i, and Δt_i is the transition time the user traveled from cluster a_i to cluster a_{i+1}.

These two sequences are considered similar if and only if they satisfy two following conditions:

- For $\forall\ 1 \leq i \leq m$, $a_i = b_i$, meaning nodes a_i and b_i must share the same cluster ID.
- For $\forall\ 1 \leq i \leq m$, $|\Delta t_i - \Delta t_i'| \leq t_{th}$, where t_{th} is a pre-defined time threshold.

Under these two conditions, a common similar sequence ($sseq$) for the two users is defined as follows:

$$sseq =< b_1(min(k_1, k_1')) \rightarrow b_2(min(k_2, k_2')) \rightarrow ...b_m(min(k_m, k_m')) >$$

m is therefore called the length of the common similar sequence of the two users.

4.3.2 Similarity measurement

The similarity of the two sequences $seq1$ and $seq2$ is determined by the measurement or the score of their common similar sequence ($sseq$), which depends on its length m:

$$s_{(m)} = \alpha_{(m)} \sum_{i=1}^{m} min(k_i, k_i') \tag{3}$$

where $\alpha_{(m)}$ is an m-dependent coefficient, for which the optimal value is determined experimentally to be $\alpha_{(m)} = 2^{m-1}$.

At a single layer of the hierarchy, two users may have multiple, say n, common similar sequences. Their similarity at a single layer is determined by the following equation:

$$S_l = \frac{1}{N_1 \times N_2} \sum_{i=1}^{n} s_i \tag{4}$$

where N_1 and N_2 denote the numbers of stay points of the two users at the layer.

The overall similarity of the two users' trajectories across the set H of multiple layers of the hierarchy is:

$$S = \sum_{l=1}^{H} \beta_l S_l \tag{5}$$

β_l is a layer-dependent coefficient, determined experimentally to be $\beta_l = 2^{l-1}$.

The advantages of this model include the exhaustive representation of users' location histories as trajectories at very fine levels of geo-spatial granularity, and the similarity measurement of two trajectories as a quantity or strength of a friendship. The model clearly shows a high level of correlation between the human movements in the real world and their social relationships. One of the disadvantages of the model is the high complexity of constructing the users' trajectories. Another disadvantage is that the issue related to coincidences was not addressed, that is when two users happen to be in the same location by accident, and possibly on multiple occasions, such as in a crowded shopping mall. Such coincidences can contribute to the similarity between two trajectories of two unrelated users and may cause a misunderstanding of a social tie existing between them.

4.4 Feature-based model

Cranshaw et al. introduced various features extracted from spatiotemporal data that have connections with, or are indications of friendships, and thus inferred friendships based on such features [3]. Similar to the approach by Crandall et al. [4] presented in Section 4.2, in order to find the co-occurrences between people, the authors first divided the space into a grid-like cells, each is of approximately 30m each side. Two people are said to co-occur if they are present in the same cell within a time interval of 10 minutes. Various features of the co-occurrences between two people were introduced, which we summarize below.

4.4.1 Diversity of a location

The primary goal of studying the diversity of a location is to evaluate the impact of a co-occurrence between two people (aka. co-location) on the fact that whether they are friends or not. Specifically, the authors aimed to find out, whether a co-occurrence between two people happened by chance (aka a coincidence) or it happened as the result of a social connection between them. For example, the fact that two people shop at the same popular mall or dine at the same popular restaurant during the same time may happen by chance, and thus they are strangers to each other. On the other hand, co-occurrences between two people at a small place, where there are only a few people, are likely a good indication of a friendship. Thus, the popularity of the location of co-occurrences matters to the prediction of friendships. Three measures are introduced to measure the diversity of a location.

Frequency is the raw number of visits by people to the location. Obviously, the higher the frequency, the more popular the location is. *User-count* is the number of unique people who have visited the location.

Location Entropy measures the diversity of a location by taking into account both the number of of unique visitors to the location, and the relative proportions of their visits. Specifically, let l be a location, let $V_{l,u} = \{< u, l, t >: \forall t\}$ be the set of visits (aka check-ins or spatiotemporal records) in location l by user u, let $V_l = \{< u, l, t >: \forall t, \forall u\}$ be the set of all visits in location l by 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:

$$H_l = - \sum_{u, P_{u,l} \neq 0} P_{u,l} \log P_{u,l} \tag{6}$$

This is called Location Entropy. 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 less popular place with few visitors, e.g., domestic houses, which are often non-crowded.

4.4.2 Features of co-occurrences

Various features of co-occurrences are introduced in this work. *Intensity* and *duration* features measure how actively (frequently) two users co-occurred, and for how long. *Location diversity* introduced in Section 4.4.1 is characterized by three different measures: frequency, user-count and location entropy. These provide the basis for understanding the impact of co-occurrences on the friendship information. *Specificity* measures how specific a location is to the user pair who co-occurred in the location. This feature is inspired by *tf-idf* [3]; specifically, the specificity of a location to the user pair u_1 and u_2 is defined as the number of co-occurrences between them in the location divided by the total number of visits by all users in the location. Some other features are related to the structural properties, such as (a) the number of people who have co-occurred with both users, (b) that number divided by the number of people who co-occurred with either user, and (c) the total number of unique locations visited by both users together divided by the total number of unique locations visited by either of the users. In addition, the regularity of each user's routine was also measured, the details of which can be found in [3].

4.4.3 The inference of friendship information

As a final step, the above-mentioned features, together with the explicit friendships (the ground truth) are used to train different classifiers, including Random Forest, AdaBoost and Support-Vector machine. The experimental results in the study [3] showed that the Random Forest and AdaBoost classifiers outperformed the Support-Vector machine classifiers.

The advantages of this model include the consideration of the popularity of the locations of co-occurrences, its impact on friendship information, and the consideration of various features of co-occurrences in inferring friendships. There are two main disadvantages of this model. First, the model only infers the binary information of friendships, meaning whether two users are friends or not, but not the strength of a friendship as compared to the two models discussed in Sections 4.2 and 4.3. Second, the use of many features may lead to the difficulty of balancing their relative importance during the training of the classifiers.

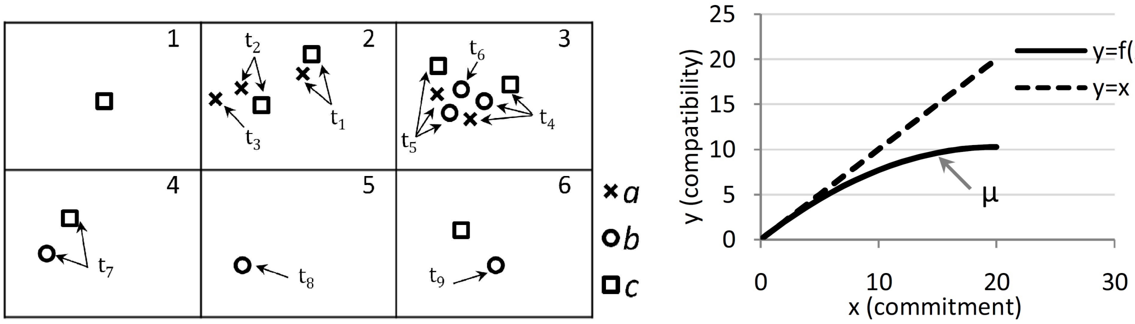
4.5 GEOSO model

In this model, Pham et al. took an entirely different approach to infer social connections from spatiotemporal data by trying to estimate the strength of people's relationships (aka ***social strength***) based on the geometric similarity of their visit patterns (i.e., who has been where and when) [5]. The authors introduced two properties: *commitment* and *compatibility*, which must be considered by any distance measure in order to correctly infer social strength from people's location behaviors.

Commitment is a phenomenon when two people repeatedly co-occurred at the same place on multiple occasions; the level of commitment is the number of times they co-occurred at the place. On the other hand, **compatibility** is a phenomenon when two people co-occurred at multiple different places; we say that two people are compatible to each other because they share a variety of common interests, which, in this case, are the places they co-visited. The main question is which, commitment or compatibility, is a better indication of a friendship? Intuitively, two close friends tend of hang out together in many different locations, and thus should co-occur in various places. On the other hand, if two people co-occurred frequently, but at only one place, they may or may not be friends because their co-occurrences may be coincidences. Therefore, the intuition is that compatibility should have more impact on social strength than commitment. We will see how GEOSO (standing for geo-social) model addresses this issue.

4.5.1 Data representation

Visit vector is a data structure that records the movement history of a user, specifically by indicating what places a person visited in the past, and at what time. To achieve this, the authors also divided the space into grid-like cells (see Section 4.2), where each cell has a unique ID. The grid is considered as a matrix, which is flattened into a vector by traveling from left to right and from top to bottom (row-first order). Correspondingly, for a given



(a)

(b)

Figure 2: (a) Visit history of three users a, b and c. An arrow indicates the time that a user visited the cell. (b) Commitment vs. compatibility.

user, each dimension of the visit vector represents one cell of the grid, and the value of the dimension is a list of time showing when the user visited the cell.

For users a and b in Figure 2(a), their visit vectors are following:

$$V_a = (0, < t_1, t_2, t_3 >, < t_4, t_5 >, 0, 0, 0)$$

$$V_b = (0, 0, < t_4, t_5, t_6 >, t_7, t_8, t_9)$$

Co-occurrence vector: Two users are said to co-occur (or to have a co-occurrence) if they were present in the same cell within a time interval τ (a threshold that can be taken as 30 minutes). Correspondingly, a *co-occurrence vector* is a data structure that indicates how many times two users co-occurred, and where they co-occurred. For example, the co-occurrence vector between users a and c is $C_{ac} = (0, 2, 2, 0, 0, 0)$. The formal co-occurrence vector for any two users i and j has the following format:

$$C_{ij} = (c_{i1,j1}, c_{i2,j2}, ..., c_{ik,jk}, ..., c_{iN,jN}) \tag{7}$$

where $c_{ik,jk}$, called ***local frequency***, denotes the number of co-occurrences between users i and j at cell of ID k.

Next, imagine there are two users $\hat{\imath}$ and $\hat{\jmath}$, who co-occurred more frequently than any other user pairs, both in the number of times they co-occurred in each cell, and in the number of unique cells, in which they co-occurred. Undoubtedly, this user pair would represent the strongest possible social connection among all the user pairs, assuming that the social strength is derived from co-occurrences only. We call the co-occurrence vector of $\hat{\imath}$ and $\hat{\jmath}$ the **Optimal Vector** (or the Master Vector), which is defined as follows:

$$M = (m, m, ..., m), \qquad m = max\{c_{ik,jk}\}, \quad \forall i, j, k \tag{8}$$

m is the maximum local frequency among all the user pairs in all locations. Note that M is a conceptual co-occurrence vector; there may or may not exist a user pair with the co-occurrence vector M. The useful information we obtain from M is that its length indicates the maximum possible commitment, while its direction corresponds to the maximum possible compatibility.

4.5.2 GEOSO distance measure

The social distance d_{ij} between users i and j is defined by the Pure Euclidean Distance (PED) between the co-occurrence vector c_{ij} and the optimal vector M. The shorter the PED distance, the closer c_{ij} and M are (in both direction and length), and thus the stronger the social connection. Social strength s_{ij} is therefore defined as the inverse of the distance metric d_{ij}.

$$d_{ij} = \sqrt{\sum_k (c_{ik,jk} - m)^2}, \quad s_{ij} = \frac{1}{(d_{ij} + 1)} \tag{9}$$

In the denominator of the formula of s_{ij}, 1 is added to d_{ij} to make sure that s_{ij} does not become infinity when $d_{ij} = 0$. This also normalizes the value of social strength s_{ij} to the range $[0, 1]$.

4.5.3 Commitment versus Compatibility

The remaining task is to analyze the relative importance of commitment and compatibility using the social strength define by the GEOSO model. Assume that users i and j have only x co-occurrences in one cell (say cell 1), user p and q have only y co-occurrences, all of which took place in different cells; without loss of generality, we can assume that y co-occurrences took place in the first y cells. The co-occurrence vectors are: $c_{ij} = (x, 0, ..., 0)$, $c_{pq} = (1, 1, ..., 1, 0, ...0)$. Clearly, users i and j have pure commitment, while users p and q have pure compatibility. We are interested in knowing how much of x would be equivalent to y in the sense that they both create the same social strength? To achieve this, we equalize the social strengths $s_{ij} = s_{pq}$ in order to find the equivalence relationship between x (commitment) and y (compatibility). From equation $s_{ij} = s_{pq}$, it is not difficult to find that $y = (2mx - x^2)/(2m - 1)$. Figure 2(b) shows this relationship. It is clear that compatibility is more important than commitment as it has more impact on social strength. For example, $x = 20$ would be equivalent to $y = 10$ for both to produce the same value of social strength. This observation is consistent with our intuition as multiple co-occurrences in a single location might just be an indicator of coincidences, such as students study in the same library while they are not friends, and therefore should be limited in contributing to social strength. On the other hand, co-occurrences in multiple locations are seldom coincidences and therefore should have more impact on social strength.

GEOSO model is particularly interesting for introducing the two properties of co-occurrences: commitment and compatibility, and for evaluating their relative importance or impact on social strength. The geometric social distance is intuitive and creates a quantitative value for social strength instead of just indicating the binary information of friendships. The disadvantage is that all locations are considered equally important, meaning a co-occurrence in a private office can have the same impact on social strength as a co-occurrence in a crowded cafe or mall. This same problem also occurs in the models in Sections 4.2 and 4.3.

4.6 EBM model

By proposing the EBM (Entropy-Based Model) model to infer social strength from spatiotemporal data [6], the goal of the authors is to address all the issues that were unsolved or partially addressed in the former studies. These issues include (a) quantifying social connections, (b) discounting the impact of coincidences, (c) evaluating the impact of each co-occurrence depending on its location, (c) addressing the problem of data-sparseness, and (d) improving the efficiency.

The EBM model explores two independent ways: ***diversity*** and ***weighted frequency***, through which co-occurrences contribute to social strength. Specifically, diversity measures how diverse the co-occurrences between two people are in terms of locations, while weighted frequency measures the impact of each co-occurrence individually depending on the popularity of the location of the co-occurrence.

4.6.1 Diversity of co-occurrences

Consider the co-occurrence vectors for 3 different pairs of users:

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

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 5 different locations, while in the former case the co-occurrences happened in just 3 different locations. Similarly, User 1 and User 3 co-occurred only in 2 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 [4] [3] [1]. This intuition is captured as *how diverse* their co-occurrences are. Below is the definition of diversity of co-occurrences [6]:

Definition 1: 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.

The goal is to formulate the diversity of co-occurrences by using either Shannon entropy or Renyi entropy. First, let's define some notations.

Let $r_{i,j}^{l,t} =< i, j, l, t >$ be a co-occurrence of User i and User j in location l and at time t. Let $R_{ij}^{l} = \bigcup_t r_{i,j}^{l,t}$ be the set of co-occurrences of User i and User j, which happened in location l. R_{ij} is the set of all co-occurrences of User i and User j in 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 in location l is $P_{ij}^{l} = |R_{ij}^{l}|/|R_{ij}|$. 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*):

$$H_{ij}^{S} = -\sum_{l} P_{ij}^{l} \log P_{ij}^{l} = -\sum_{l, c_{ij,l} \neq 0} \frac{c_{ij,l}}{f_{ij}} \log \frac{c_{ij,l}}{f_{ij}} \tag{10}$$

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.

Similarly, the uncertainty can also be expressed using Renyi entropy - a more generalized type of entropy with a flexibility to control the contribution of each component P_{ij}^{l}.

$$H_{ij}^{R} = \left(-\log \sum_{l} \left(P_{ij}^{l} \right)^{q} \right) / (q-1) \tag{11}$$

$$= \left(-\log \sum_{l} \left(\frac{c_{ij,l}}{f_{ij}} \right)^{q} \right) / (q-1) \tag{12}$$

where $q \geq 0$ is the order of diversity.

Generally, entropy is often regarded to as the *index* of diversity, but not diversity itself [8]. Diversity D is computed as the exponential function of entropy H. Specifically, $D = \exp(H)$. The expressions for diversity using each of the entropies above is:

$$D_{ij}^{S} = \exp \left(-\sum_{l, c_{ij,l} \neq 0} \frac{c_{ij,l}}{f_{ij}} \log \frac{c_{ij,l}}{f_{ij}} \right) \tag{13}$$

$$D_{ij}^{R} = \left(\sum_{l, c_{ij,l} \neq 0} \left(\frac{c_{ij,l}}{f_{ij}} \right)^{q} \right)^{1/(1-q)} \tag{14}$$

The upper index S denotes *Shannon*, R denotes *Renyi*.

Both Shannon entropy and Renyi entropy show how diverse a co-occurrence vector is in terms of locations. It is the *unpredictability* of the location of a co-occurrence. In other words, it is the amount of location information in the co-occurrences of two users. Therefore, their advantage is that its capture of diversity is consistent with the intuitions of friendships. First, the more locations, the higher the entropy. This is intuitive as the more places two users visited together, the stronger their connection. Second, the more uniform the distribution of the co-occurrences across locations (more equal proportion of co-occurrences in each location), 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.

However, the disadvantage of Shannon entropy is that it may give higher importance to large components (aka outliers) of the co-occurrence vector because each component is weighted by its proportional abundance. For example, 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. This is not always a desired behavior that we want, because a high number of co-occurrences in a single crowded location may indicate coincidences, and their contribution to the social strength should, in fact, be limited rather than amplified.

On the other hand, Renyi entropy can effectively address the problem of coincidences. 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 $c_{ij,l}$. Specifically:

- 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 impact the outliers have on Renyi entropy.
- When $q < 1$, instead, the Renyi entropy gives more weight to the *low* local frequencies $c_{ij,l}$.
- 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*.
- When $q = 1$: 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 *cross-over* point where Renyi entropy stops all of its biases and weighs the local frequencies $c_{ij,l}$ by their *own* relative proportions, which is what Shannon entropy does. Thus, at $q = 1$, Renyi entropy becomes Shannon entropy. Indeed, even though Equations (11) and (12) are *undefined* at $q = 1$, their limits exist when $q \to 1$ and become the Shannon entropy.

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. The readers are referred to [6] for how to obtain the optimal order of diversity experimentally.

4.6.2 Weighted frequency

While diversity measures the *breadth* of co-occurrences across locations, weighted frequency, on the other hand, measures the *depth* of co-occurrences and weighs each co-occurrence individually depending on the popularity of the location. Weighed frequency utilizes Location Entropy, which was discussed in Section 4.4.1. The formula of weighted frequency is given as follows:

$$F_{ij} = \sum_{l} c_{ij,l} \times \exp(-H_l) \tag{15}$$

Weighted frequency tells us how important the co-occurrences at non-crowded places are to social connections. Crowed locations have high Location entropy H_l, thus low $\exp(-H_l)$, and consequently the impact of $c_{ij,l}$ on F_{ij} is decreased. On the other hand, for non-crowded locations, $\exp(-H_l)$ is high and this increases the impact of $c_{ij,l}$. The authors also provided more details about weighted frequency, including its comparison to *tf-idf*, and how weighted frequency addresses the problem of data sparseness [6].

4.6.3 Social strength

Finally, diversity and weighted frequency are combined to create social strength. Let s_{ij} be the ultimate social strength that captures both diversity and weighted frequency. A linear regression is conducted:

$$s_{ij} = \alpha.D_{ij} + \beta.F_{ij} + \gamma \tag{16}$$

where D_{ij} and F_{ij} are defined in Equations (14) and (15), respectively. Parameters α, β and γ can be either learned from dataset, or provided by users, or provided as application-dependent parameters. As a good practice, s_{ij} is generally normalized to $[0, 1]$. The information about how to obtain the parameters of the regression can be found in [6].

The advantages of the EBM model include the capture of the intuition of social connections in co-occurrences, specifically by measuring the diversity of co-occurrences in terms of locations and the weighted frequency. While the diversity offers a flexible mechanism of eliminating the impact of coincidences through Renyi entropy, weighted frequency takes into account the impact of each individual location of co-occurrences by analyzing the popularity of each location through Location Entropy. In general, all the main concerns of the former models we pointed out in the previous sections have been effectively addressed by the EBM model.

5 Conclusion

In this article, we surveyed the solutions proposed for inferring the real-world social connections from spatiotemporal data. Toward this end, we presented various models in details; for each model, we discussed the key ideas/intuitions of how social connections are linked to the location history of users. We also explained the main formulations of social connections and social strength for each model, together with its advantages and disadvantages.

This line of research opens a number of opportunities for future work. For example, the inferred real-world social connections and their strengths can be used to further study other aspects of social networks, such as social influence and information propagation among people in the real world. It is also possible to investigate the type of each social connection, whether two people are in a casual friendship, colleagues or in a family relationship, based on the semantics of the locations, in which they co-occurred. The real-world social connections can also be applied in other fields of study, such as in epidemiology to study the spread of disease through human contacts, or in criminology to investigate the nature, causes, patterns and consequences of a criminal behavior.

6 Acknowledgements

This research has been funded in part by NSF grants IIS-1115153, IIS-1320149, and CNS-1461963, the USC Integrated Media Systems Center (IMSC), and unrestricted cash gifts from Google, Northrop Grumman, Microsoft, and Oracle. 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.

References

[1] Eagle, Nathan, Alex Sandy Pentland, and David Lazer. “Inferring friendship network structure by using mobile phone data.” Proc. of the National Academy of Sciences 106, no. 36 (2009): 15274-15278.

[2] Li, Quannan, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. “Mining user similarity based on location history.” Proc. of the 16th ACM SIGSPATIAL, p. 34. ACM, 2008.

[3] Cranshaw, Justin, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. “Bridging the gap between physical location and online social networks.” Proc. of the 12th ACM Ubicomp, pp. 119-128. ACM, 2010.

[4] Crandall, David J., Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. “Inferring social ties from geographic coincidences.” Proc. of NAS 107, 22436-22441, 2010.

[5] Pham, Huy, Ling Hu, and Cyrus Shahabi. “Towards integrating real-world spatiotemporal data with social networks.” Proc. of the 19th ACM SIGSPATIAL, pp. 453-457. ACM, 2011.

[6] Pham, Huy, Cyrus Shahabi, and Yan Liu. “Ebm: an entropy-based model to infer social strength from spatiotemporal data.” Proc. of the 2013 ACM SIGMOD, pp. 265-276. ACM, 2013.

[7] http://mashable.com/2012/01/09/real-world-digital-world/

[8] Jost, Lou. “Entropy and diversity.” Oikos 113, no. 2 (2006): 363-375.

[9] Kempe, David, Jon Kleinberg, and Eva Tardos. “Maximizing the spread of influence through a social network.” Proc. of the ninth ACM SIGKDD, pp. 137-146. ACM, 2003.

[10] Leskovec, Jure, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. “Cost-effective outbreak detection in networks.” Proc. of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 420-429. ACM, 2007.

[11] Chen, Wei, Laks VS Lakshmanan, and Carlos Castillo. “Information and influence propagation in social networks.” Synthesis Lectures on Data Management 5, no. 4 (2013): 1-177.

[12] Rogers, Everett M., and F. Floyd Shoemaker. “Communication of Innovations; A Cross-Cultural Approach.” (1971).

[13] LibenNowell, David, and Jon Kleinberg. “The linkprediction problem for social networks.” Journal of the American society for information science and technology 58, no. 7 (2007): 1019-1031.

[14] Ugander, Johan, Brian Karrer, Lars Backstrom, and Cameron Marlow. “The anatomy of the facebook social graph.” arXiv preprint arXiv:1111.4503 (2011).

[15] Wu, Peng, and Dan Tretter. “Close and closer: social cluster and closeness from photo collections.” Proc. of the 17th ACM international conference on Multimedia, pp. 709-712. ACM, 2009

[16] Noulas, Anastasios, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. “An Empirical Study of Geographic User Activity Patterns in Foursquare.” ICwSM 11 (2011): 70-573.

[17] Roussopoulos, Nick, Stephen Kelley, and Frdric Vincent. “Nearest neighbor queries.” In ACM sigmod record, vol. 24, no. 2, pp. 71-79. ACM, 1995.

[18] Sharifzadeh, Mehdi, and Cyrus Shahabi. “The spatial skyline queries.” Proc. of the 32nd international conference on Very large data bases, pp. 751-762. VLDB Endowment, 2006.