GEOSO – A Geo-Social Model: From Real-World Co-occurrences to Social Connections[1]

Huy Pham, Ling Hu, Cyrus Shahabi

Integrated Media Systems Center
University of Southern California (USC)
Los Angeles, California, USA
{huyvpham, lingh, shahabi}@usc.edu

Abstract. As the popularity of social networks is continuously growing, collected data about online social activities is becoming an important asset enabling many applications such as target advertising, sale promotions, and marketing campaigns. Although most social interactions are recorded through online activities, we believe that social experiences taking place offline in the real physical world are equally if not more important. This paper introduces a geo-social model that derives social activities from the history of people's movements in the real world, i.e., who has been where and when. In particular, from spatiotemporal histories, we infer real-world co-occurrences - being there at the same time - and then use co-occurrences to quantify social distances between any two persons. We show that straightforward approaches either do not scale or may overestimate the strength of social connections by giving too much weight to coincidences. The experiments show that our model well captures social relationships between people, even on partially available data.

Keywords: Data mining, geospatial, spatiotemporal, social network

1 Introduction

Nowadays, a significant amount of social interactions are gathered from various online activities of Internet users. These virtual social events provide important cues for inferring social relationships, which in turn can be used for target advertising, recommendations, search customization, etc., the main business model of Internet giants. However, an important aspect of the social network is overlooked – the fact that people play active social roles in the physical world in their daily lives. However, as most social interactions and events that take place in the physical world are not as well documented as the ones that can be acquired from an online social network application, it is necessary to seek for alternative methods to infer social relationships from people's behavior in the physical world.

With the popularity of GPS-enabled mobile phones, cameras, and other portable devices, a large amount of spatiotemporal data can easily be collected or is already

[1]This paper is a full version of a poster paper appeared in ACMGIS'2011 [19].

available. Those data in their simplest form capture people's visit patterns, i.e., *who has been where and when*. However, we believe that the information hidden behind those data is a strong indicator of the social connections between people in their real lives [5,6]. Intuitively speaking, if two people happen to be at the same place around the same time for multiple occasions, it is very likely that they are socially involved in some way.

One area of related work includes a number of recent studies [20,21,22,23,24] that investigate similarity between objects' (e.g., people, cars) locations in time, represented as trajectories, for various reasons (e.g., to identify moving convoys, to recommend carpool partnerships). In particular, the similarity between trajectories can be used to infer social connections among people as shown by Li et al. [3]. The concept of trajectories in these studies usually indicates a shorter duration of time (in the order of hours) in which the sequence/order of visits is important. However, in our case co-occurrences refer to longer-term shared locations (in the order of months) in which the sequence of visits has no significance. Hence our "vector" model that captures "frequency" of co-visits and our various distance measures (used between the vectors) are very different than those suggested by these related studies.

One of the few papers that study the inference of social connections from real-world co-occurrences is by Crandall *et al.* [1]. They applied a probabilistic model to infer the probability that two people have a social connection, given that they co-occurred in space and time. Their method not only takes into account both spatial and temporal features but also handles sparsely distributed spatial data. However, they do not consider the frequency of co-occurrences in space and time, which we argue that it is an important indicator of social connections. Moreover, to render the problem tractable, Crandall *et al.* made the simplifying assumption that each person has one and only one friend, generating a sparse graph of M vertices and M/2 edges, where M is the total number of the users. Unfortunately, this assumption may not hold in many cases, as the social connection network can be quite dense in real world. For example, consider a group of people, who work for the same company; they are all socially connected to each other as co-workers.

In this paper, we take an entirely different approach to this problem by trying to estimate the strength of people's relationships based on the similarity of their visit patterns (i.e., *who has been where and when).* Hence, the questions we focus on are how to represent people's visit patterns (in space and time) and how to measure the distance between these visit patterns.

One intuitive solution is to represent the visit patterns as time-series (by transforming 2-D space to 1-D location ID's on the y-axis), and then apply a cross-correlation integral to measure the similarity between two time-series of two users. Besides the fact that this approach would not scale due to the eternity of time, in Section 2.2, we show that a more important problem with this approach is that the y-axis (representing the 2D space) of the time-series reflects a false notion of continuity of space, resulting in misrepresentation of the visit information in time intervals between two visits.

Alternatively, a person's visit pattern can be modeled as a vector where each dimension corresponds to a fixed location ID (again, by transforming 2-D space to 1-

D), and the values capture the frequency of visits. Consequently, we can apply distance metrics, such as the cosine similarity [2] to calculate the distance between two patterns represented by vectors. However, we show in Section 2.2 that there are two major drawbacks with this approach. That is, it does not preserve the temporal feature and it cannot differentiate a vector $\vec{v}$ with its scaled counterpart $k\vec{v}$, both of which are crucial to our problem.

Since straightforward representations and distance measures do not work, in this paper, we propose a new representation along with a corresponding distance measure. In addition, and more importantly, we identify two properties, *commitment* and *compatibility*, that any distance measure should have in order to correctly infer social strengths from co-occurrences. We call this collection of contributions as a new model, dubbed Geospatial Social Model (*GEOSO*), towards integrating real-world spatiotemporal data with social-networks.

Our representation of visit patterns is a slight modification of the vector representation with time information captured at each dimension of the vector. However, we show in Section 4.3 that the simple cosine or Euclidean distance measures on this new representation cannot capture both of our properties, resulting in wrong estimation of social connectivity. Therefore, we discuss various auxiliary representations such as *co-occurrence vector* and *master vector*, to enable an accurate distance computation.

We experimentally evaluate the *GEOSO* distance model using data from the Internet Movie Database-IMDB (for co-occurrence events and social connections) and Wikipedia (for social connections). We compute the social distances based on co-occurrence events of celebrities, and validate the results with the social connection information available from their Bio on IMDB and Wikipedia. That is to verify whether user pairs with small social distances in our model have close social relationships in reality, e.g., close friends, siblings, life partners, etc. Our experiments show that the precision of our distance model is over 80% for user pairs with distance values less than 0.5.

The remainder of this paper is organized as follows. Section 2 formally defines the problem and shows why existing similarity metrics do not apply to our problem. In Section 3, we introduce the *GEOSO* model which quantifies the social distances between user pairs. In Section 4, we prove that GEOSO captures our two social properties. We validate our model through extensive experiments and report the results in Section 5. Finally, we conclude the paper with future directions in Section 6.

2 Problem Definition

2.1 The Problem

Given a set of users $U = (u_1, u_2, \dots, u_M)$, a set of places $P = (p_1, p_2, \dots p_N)$, and a set of spatiotemporal social events, the problem is how to infer the social connections between each pair of users and how to measure the social connections based on certain quantitative values. As part of the input data, social events are represented by a set of triplets $< u, p, t >$ stating *who (u) visited where (p) and when (t).* The

temporal feature of the event can be either a time-stamp or a time interval, whichever is available. We term the event triplets as W^3 events.

Intuitively speaking, people who are socially close to each other have higher chances of visiting same places at the same time (co-occurrences in both space and time). For example, a couple who lives together probably visits same grocery shops, restaurants, and vacation destinations at the same time. Furthermore, people who repeatedly visit the same location at the same time are socially connected with higher probability. For example, co-workers go to work on every weekday. Subsequently, we declare the following observations for the ease of discussion and refer to them later.

Observation 1 The more places two users visited together at the same time, the more likely these two users are socially close to each other.

Observation 2 The more often two users visited same places at the same time, the closer the two users are socially connected.

2.2 Candidate Similarity Metrics

As discussed earlier, the W^3 event history of any person can be easily represented as a vector or a time-series. Therefore, applying existing similarity metrics to our problem appears to be promising. In this section, we discuss two existing similarity metrics and point out why these candidate solutions do not apply to our problem.

2.2.1 Cross-Correlation Integral

Cross-correlation integral is frequently used in signal processing [4,7] to measure the similarity of two waveforms as a function of time. It also applies to pattern recognition problems [8,15] to find the similarity between two patterns. We can use cross-correlation integral to measure the visit patterns of two users in space and time. Particularly, let the x-axis be time and y-axis the geo-spatial locations, e.g., the label of grid cells if we consider the whole 2D space as a grid and number the cells in row-order. Each W^3 event corresponds to a point in the coordinate system and points are connected chronically using linear interpolation. Consequently, we have one time-series for each user as shown in Fig. 1. Next we compute the cross-correlation integral based on the time-series of two users and use the result as the similarity measure of the two users.

However, there are two major problems with this approach when applied to our problem. First, as the time-series is a function of time, it does not scale well. When the time axis is continuously growing, it results in a linear increase in time complexity of any possible similarity function. This shows that representing user visit patterns as a function of time and space is not appropriate for our problem. Second, the space is discretized as non-overlapping cells and the cells on the y-axis may be numbered in an arbitrary way (in row order or Hilbert curve order). Thus, being in two cells, for example, cell *x* and cell *z*, at two time instances does not indicate that the user was ever in any intermediate cells that lie spatially between cell *x* and cell *z*. Therefore, the time-series can misinterpret the visit pattern of the user, and the cross-correlation

integral over time-series of two users may lead to imprecise results and hence incorrect social distance measurements.

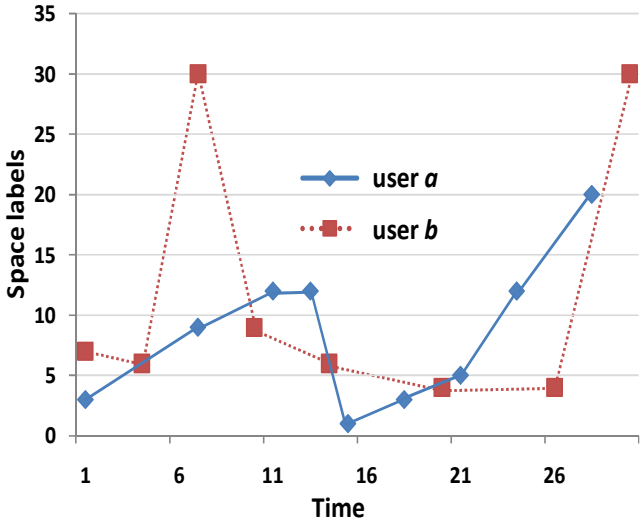

Fig. 1. Cross-correlation integral of two user visit patterns

2.2.2 Cosine Similarity

Cosine similarity measures the similarity between two vectors based on the cosine value of the angle between them. In the field of information retrieval, cosine similarity is often used to compare the similarity between two documents [9,10]. If we consider the user visit patterns as vectors, the cosine similarity metric can be adopted to solve our problem. Let V_a be a vector which records the number of times that user a visits a geo-location in the space and V_b be the same vector for user b. We can compute the cosine similarity between the two vectors V_a and V_b, which is then used to measure the social distance between a and b. However, there is a major drawback in this approach because the time dimension is overlooked in the vector representation of the visit history. For example, if both user a and user b have visited the same geo-locations, but on different days, they are considered similar in this approach but they are not similar in reality as they have never been at the same place *at the same time*. Obviously, the simple vector representation cannot handle the time dimension, which is an important factor in measuring social distances.

Furthermore, cosine similarity essentially measures the cosine of the angle between two vectors, therefore, the scalar of the vectors are not measured or considered. That is, the cosine similarity between a vector $\vec{v}$ and $\vec{u}$ is the same as the cosine similarity between $k\vec{v}$ and $\vec{u}$. This is not appropriate in measuring social distances based on visit patterns as the number of visits is an important indication of social closeness.

3 The GEOSO model

To better capture the relationship between spatiotemporal co-occurrences and social ties among people, we propose a geo-social data model, called *GEOSO*.

3.1 Data Representation

Assume that the data input to the problem is a sequence of triplets in the form of <***user, location, time***>, specifying who visited where and when. Following the storage model in [1,11], the 2D space, formed by latitude and longitude, is partitioned into disjoint cells. For example, the space could be divided by a grid consisting of X × Y rectangular cells. The size of the cells is application-dependent and we discuss it later in the experiment section.

3.1.1 Visit Vector

A **visit vector** is a data structure that records the movement history of a user. We consider the grid as a matrix and then store it in row-first order as a vector. Subsequently, for each user, a **visit vector** is constructed to record the visit history of that user within a period of time. Specifically, 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 these visits to the cell happened. If the user has not visited a cell within the time period of interest, the value of that cell is 0. For example, in Fig. 2, the visit vectors of user *a* and user *b* are:

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

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

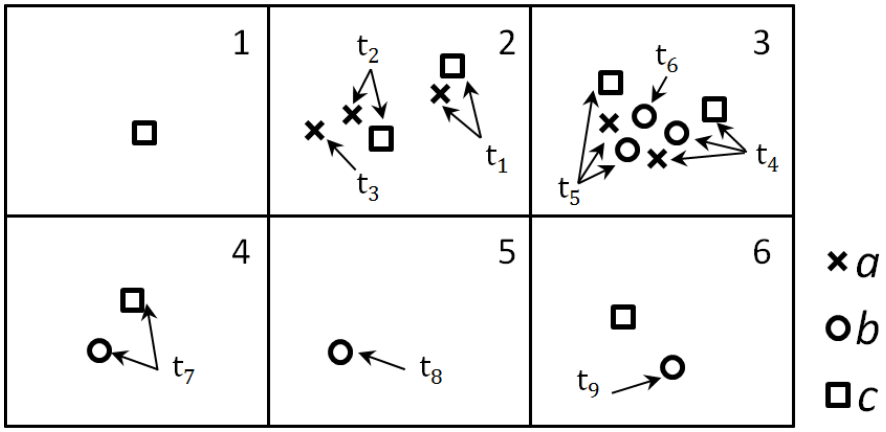

Fig. 2. Visit history of user a, b and c

As most users only visit a fairly small number of cells compared to the total number of cells in the space, the visit vector for a single user may contain mostly zeros and a few non-zero values. For storage and computation efficiency, we can eliminate all zeros and only store the non-zero values together with their cell IDs. For example, the visit vector of user *a* in Fig. 2 can be stored as $V_a = (2{:}<t_1, t_2, t_3>, 3{:}<t_4, t_5>)$, which represents that user *a* visited cell *2* for three times and cell *3* two times. For ease of presentation, we still use the original representation of the visit vector throughout the rest of the paper but keep in mind that the vectors can be stored and computed in a more efficient way.

3.1.2 Co-occurrence Vector

Next, we define a data representation to capture the commonalities between two users. The **co-occurrence vector** states the common visits of two users for the time period of interest. Each dimension of the vector still corresponds to a cell in the grid. However, the value of each dimension does not record the time of the visits but the number of times that the two users visited the same cell at *roughly the same time*, that is, the time spans of the visits of the two users at the same cell overlap. Note the length of the time overlap is application dependent and can be an input parameter to our model, for example, 20 minutes or two hours. Consider users *a* and *c* in Fig. 2, both *a* and *c* visited cells *2* and *3* at the same time. In particular, *a* and *c* visited cell *2* two times and cell *3* two times together. The co-occurrence vector between user *a* and *c* is $C_{ac} = (0,2,2,0,0,0)$. We formally define the co-occurrence vector as follows:

$$C_{ij} = (c_{i1,j1}, c_{i2,j2}, \dots, c_{iN,jN}) \quad (1)$$

In Eq. 1, a term $c_{ik,jk}$ denotes the number of times that user *i* and user *j* both visited cell *k* while *k* ranges from *1* to the total number of cells *N*. Note that co-occurrence vectors can also be stored in a compact form, while only non-zero values are stored and maintained. In the next section, we discuss how to perform computation efficiently on these compact vectors.

3.1.3 Master Vector

As two co-occurrence vectors can considerably differ from each other, we need to normalize co-occurrence vectors so that the distance measurements are comparable. Consider that two users *i* and *j* have visited every cell in the space at the same time, and the number of visits to each cell is the maximum among any pair of users in the group of users of interest. Let C_{ij} be the co-occurrence vector of *i* and *j*. Undoubtedly, user *i* and user *j* have the highest similarity, hence, the smallest distance between each other. Furthermore, the more similar the co-occurrence vectors of any user pair to C_{ij}, the closer the two users are in terms of social distance. Following this intuition, we define the **master vector** for a group of users. A master vector contains the maximum pair-wise co-occurrences in each cell for a group of users of interest. For instance, the co-occurrence vectors of users *a*, *b* and *c* in Fig. 2 are as follows:

$$C_{ab} = (0,0,2,0,0,0)$$
$$C_{ac} = (0,2,2,0,0,0)$$
$$C_{bc} = (0,0,2,1,0,1)$$

The master vector of the three users is $M = (0,2,2,1,0,1)$ where the value of each dimension of M is the maximum value of the three co-occurrence vectors at the corresponding dimension. Note that only one master vector is constructed for a given set of users. Computing the master vector is simple and can be done efficiently. The definition of the master vector is shown in Eq. 2, where *U* stands for the total number of users and *N* is the total number of cells.

$$M = (m_1, m_2, \dots, m_k, \dots, m_N) \tag{2}$$
$$m_k = \max_{1 \le i < j \le U, 1 \le k \le N} c_{ik,jk}$$

3.2 The GEOSO Distance Measure

The goal of our problem is to efficiently compute the social connections among all pairs of users and report those users who are strongly bonded. For any given set of users and their W^3 events, we first compute the co-occurrence vectors for every pair of users and the master vector for the entire set of users. Next, we compute the social distance between each pair of users.

The social distance d_{ij} between user *i* and user *j* is defined by the Pure Euclidean Distance (PED) between the co-occurrence vector C_{ij} and the master vector *M*. The similarity s_{ij} between two users is the inverse of the distance metric.

$$d_{ij} = \sqrt{\sum_k (c_{ik,jk} - m_k)^2}, \qquad s_{ij} = \frac{1}{d_{ij}} \tag{3}$$

Consider a simple example consisting of two cells and three users shown in Fig. 3. The x-axis shows the number of co-occurrences in cell *1* and the y-axis shows the number of co-occurrences in cell *2*. The co-occurrence vectors are plotted as thinner arrowed lines and the master vector is plotted with a solid bold arrowed line. The co-occurrence vector of user *a* and *b* is $(2,2)$, the co-occurrence vector of users *a* and *c* is (0,3), and the co-occurrence vector of users *b* and *c* is $(0,2)$. The master vector of the three users is $M = (2,3)$.

Next, the PED distances between all user pairs are computed as follows:

$${d_{ab}}^2 = (2-2)^2 + (3-2)^2 = 1$$
$${d_{ac}}^2 = (2-0)^2 + (3-3)^2 = 4$$
$${d_{bc}}^2 = (2-0)^2 + (3-2)^2 = 5$$

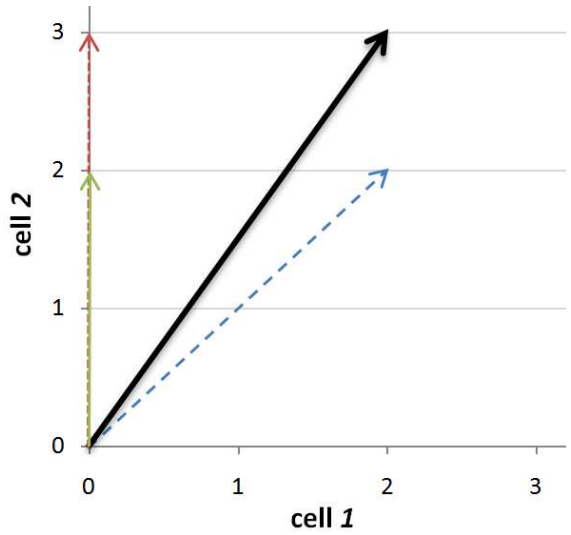

Fig. 3. Vector view of GEOSO distance measurements.

The smaller the distance between two users, the closer they are. Therefore, we know that users *a* and *b* are the closest user pair in the example shown in Fig. 3.

As co-occurrence vectors contain mostly zeros, they are stored in a compact form. That is, all zeros are eliminated from the vector. Subsequently, we can improve the computation efficiency by employing the Projected Pure Euclidean Distance (**PPED**) proposed in [2].

4 Properties of the GEOSO Measure

In this section, we introduce two important properties of the *GEOSO* model and how our model captures the properties quantitatively.

4.1 Compatibility

According to the first observation in Section 2.1, the more common cells two users visit, the higher the likelihood that these two users are socially closer. Now, we show that our social distance measure is consistent with this observation. First, let us temporarily not consider the number of co-occurrences in one cell between two users, but only the fact whether two users co-occurred in that cell. In the co-occurrence vector, if two users both visited a cell at the same time (co-occurred), we use the value *1* for that cell. Otherwise, we assign *0* to that cell. Note that the dimensionality of the vector stays the same. In the extreme case, if two users visited every cell together, their co-occurrence vector contains only ones in all dimensions. Generally,

suppose we have two pairs of users, i.e., (i, j) and (p, q). Users i and j both visited k cells together, while users p and q both visited $k + a$ cells together ($a > 0$). The co-occurrence vectors of the two user pairs are:

$$C_{ij} = (1,1, \dots, 1,0,0, , \dots \quad ,0)$$

$$C_{pq} = 1,1, \dots, 1,1, \dots 1,0, \dots, 0)$$

Without loss of generality, suppose all co-occurrences happened in the first several cells. Clearly, the social distance between the user pair (p, q) is closer because $p$and $q$has more overlap in space and time. We define the total number of dimensions with non-zero values in the co-occurrence vector as the **compatibility** between the two users. Then, compatibility property says that the more compatible two users are in their social relations, the closer they are. Next, we prove that our distance model captures the compatibility property.

Consider a new master vector that is represented as $M' = (m, m, \dots, m)$ where m is the maximum value of all dimensions in the original master vector in Eq. 2. Note that the new master vector M' changes the absolute distance values but does not change the relative values between two distances. That is, if d_{ij} is greater than d_{pq} with regard to the original master vector M, it is still greater than d_{pq} with regard to the new master vector M'. Consequently, we use M' instead of M as the master vector in the following discussions where only the relative distance values are of concern. Hence, the distances between user *i* and *j*, *p* and *q* are as follows:

$$d_{ij} = \sqrt{k(m-1)^2 + (N-k)m^2}$$

$$d_{pq} = \sqrt{(k+a)(m-1)^2 + (N-k-a)m^2}$$

Next, consider the difference between the two distances:

$$d_{ij}^2 - d_{pq}^2$$

$$= k(m-1)^2 + (N-k)m^2 - (k+a)(m-1)^2 - (N-k-a)m^2$$

$$= -a(m-1)^2 + am^2 = a(2m-1) > 0 \text{as} m > 0$$

Hence d_{ij} is greater than d_{pq}. Consequently, user p and q are more socially connected than user i and j. Note that if m equals to zero, it is a trivial case where no two users visited the same cell and their distances are all set to infinity. Therefore, our model has the compatibility property.

4.2 Commitment

As stated in our second observation, if two users repeatedly visited the same places together, they are more likely socially close to each other. For examples, the fact that students go to the same classroom twice a week is a strong indication that they are classmates. To show that our distance model is consistent with this observation, we

need to take into account the number of co-occurrences between two users which we left behind in the previous section. That is, the value of each dimension in the co-occurrence vector corresponds to how many times two users co-occurred in space and time. Then the second observation states that the more two users committed to a certain place, the closer they are. We call it the **commitment** property of social relations. Next we prove how the model captures the commitment property.

Suppose that the co-occurrence vectors of two pairs of users (i,j) and (p,q) are identical except in one dimension.

$$C_{ij} = (k, \quad c_2, c_3, \dots, c_N)$$

$$C_{pq} = (k + a, c_2, c_3, \dots, c_N) \ (a > 0)$$

The distances between the two pairs of users are:

$$d_{ij} = \sqrt{(m-k)^2 + \beta}$$

$$d_{pq} = \sqrt{(m-k-a)^2 + \beta}, \beta = \sum_{2 \le l \le N} (m - c_l)^2$$

Next, consider the difference of the two distances:

$$d_{ij}^2 - d_{pq}^2 = (m-k)^2 - \ (m-k-a)^2 > 0$$

Hence d_{ij} is greater than d_{pq}. Therefore we conclude that p and q are more socially connected than i and j. This shows that our model has the commitment property.

4.3 Compatibility vs. Commitment

We have shown that both compatibility and commitment properties play important roles in measuring social distances and they are captured by our *GEOSO* model. As the next step, we analyze the relationship between the two in the model and show which of the two properties are more important.

Assume user i and j have x co-occurrences in one cell (say cell *1*), user p and q have y co-occurrences all of which happened in different cells. Without loss of generality, suppose that y co-occurrences happened at the first y cells. The co-occurrence vectors are:

$$C_{ij} = (x, 0, 0, \quad \dots, 0)$$

$$C_{pq} = (1, 1, \dots, 1, 0 \dots, 0)$$

The distances functions are:

$$d_{ij} = \sqrt{(m-x)^2 + (N-1)m^2}$$

$$d_{pq} = \sqrt{y(m-1)^2 + (N-y)m^2}$$

Let $d_{ij} = d_{pq}$ and we have the relationship between x and y as the quadratic function shown in Eq. 3.

$$y = f(x) = \frac{2mx - x^2}{2m - 1} \tag{4}$$

In the equation above, m is a constant. The relationship between the variable x and variable y is plotted in Fig. 4 (m is set to *20*).

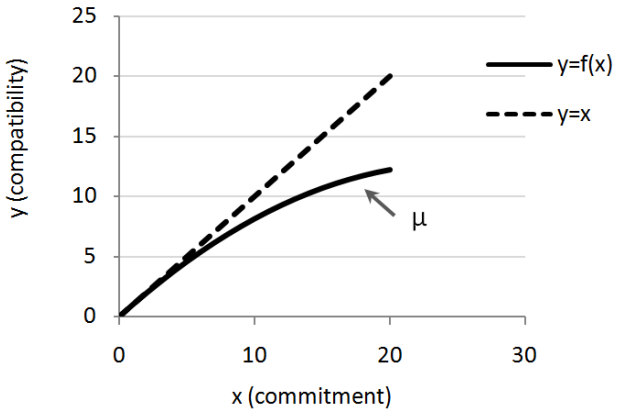

Fig. 4. Commitment vs. compatibility

The figure of the relationship between commitment and compatibility gives two important insights. First, as the curve of $y = f(x)$ is always below the line of $y = x$, our models shows that the commitment property has less importance on the distance function than the compatibility property. This is consistent with some intuitive examples. Consider the activities of two students on campus. If their W^3 event history shows that they went to the cafeteria *10* times together, the gym *10* times together and the same classroom *4* times together in the past month (high compatibility), this is a strong indication that these two students are close friends. On the other hand, if two students have been to the library at the same time for 30 times (high commitment), it does not necessarily show that the two are friends. In fact, there might be hundreds of students who go to the library every day. However, most of them do not know other students who also study in the same library.

Second, it is shown in the Fig. 4 that as commitment (x) increases, compatibility (y) also increases, however, with a much slower speed. We can increase either the commitment or the compatibility to yield a certain social distance. However, it requires less change in compatibility than commitment. When commitment reaches its upper limit (the saturation point) μ, further increasing commitment only very insignificantly affects the social distance of our model. This also confirms the fact that a spike of large commitment value only implies a coincidence in our social lives and does not bring closer the social distances.

The *GEOSO* model captures both compatibility and commitment properties of social behaviors by applying both the co-occurrence vectors and the master vector

collectively. Without these data representations, applying the simple cosine or Euclidean distance measures on the simple visit vectors of users will lead to wrong estimation of social connectivity, in particular, the commitment property will overestimate social distances and weaken the influences of compatibility. For example, two users that co-occurred in the same places together for *k* times will have the same social distance as two users that co-occurred in *k* different places but only once in each place in both cosine similarity or Euclidean distance measure.

5 Experiment

5.1 Dataset

Ideally, we want to first compute the social distances between user pairs by applying the *GEOSO* distance model to a dataset of visit patterns (*who has been where and when*), Subsequently, we compare our results with the real social distances of the same set of users and measure the *precision and recall* of results. However, data that include both spatiotemporal information and real social connections among the same set of people are often considered sensitive and private. One can easily find either a spatiotemporal dataset (e.g., extracted from photos on Flickr [12]) or a dataset with social connections (e.g., LiveJournal [13]) separately. However, to the best of our knowledge, datasets with the combination of the two are not fully available for public or research uses.

Consequently, we seek an alternative solution and decide to use data from the Internet Movie Database (IMDB) [14] because it resembles the data requirements of our experiments for two reasons. First, the dataset contains spatiotemporal data of people. For example, if two actors/actresses acted in the same movie/episode, we consider that two persons co-occurred in space and time. If they performed in more than one movie/episode, we consider that they co-occurred in space and time multiple times as an indicator of compatibility, and if they performed in multiple episodes of the same TV series, it is considered an indicator of commitment. The social distance *d* (see Eq. 3) is calculated for each pair of people. Second, social connections of these actors/actresses are available publicly. For example, the Bio sections on IMDB and/or the Wikipedia [18] web pages usually contain the social relationships of that actor/actress, such as parents, siblings, spouses, best friends, long-time acting partners, etc. These data of social connections can be used to verify if two people with short social distance *d* is indeed socially connected. One might argue that the fact of two actors/actresses performing in the same movie does not necessarily suggest that they are related. This is a valid argument. However, the same thing is also true in a real spatiotemporal dataset, that is, two persons appearing at the same place at the same time may only due to coincidences. Our model can handle these coincidences by weighting compatibility and commitment appropriately in a non-linear fashion (See Fig. 4).

We extracted the information as described above from the IMDB and Wikipedia and ran our experiments on these datasets. Table 1 provides an overview of the

datasets used in this section. The first row describes the sizes of celebrity sets, and the second row shows the number of different movies that the corresponding set of celebrities acted in. The last row of the table summarizes the total number of tuples in the format of <*person*, *movie/episode*, *time*>, which corresponds to the <who, where, when> (W^3) events.

Table 1. Dataset

# of celebrities	2k	4k	10 k
# of movies and episodes	32k	50k	100k
# of W^3 events	280k	1.1M	4.6M

5.2 Distance Measure and Result Verification

In this section, we ran experiments on each data set and computed social distances using the *GEOSO* model. Next, the distances are normalized and discretized. We divided [0, 1] into 25 of equal-sized buckets and each bucket contains user pairs with distances between a and $a + 1/25$. For example, the first bucket contains user pairs with distances between 0 and 0.04.

The first dataset contains 2,000 celebrities with 280k co-occurrences. Most of the user pairs (96.8% of 280k) have social distances close to 1 (> 0.91), meaning that they are socially far away. Therefore, we drop those user pairs and focus on those who are socially close, which is 8.8k pairs (3.2% of 280k).

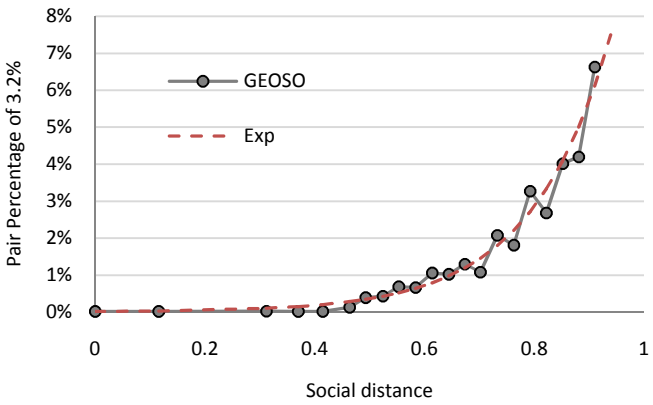

Fig. 5. Percentage of pairs vs. social distances

The relationship between distance values and the user pair percentage, which is calculated out of 8.8k pairs, is shown in Fig. 5. The x-axis shows the social distance calculated by our model and the y-axis shows the percentage of top 3.2% user pairs with smaller distances. Each tipping point in the graph represents a bucket, and as the graph shows, buckets with short social distances have fewer pairs of people (lower

percentage) than buckets with long social distances do. Keep in mind that the number of buckets (number of tipping points) does not represent the number of pairs of people, however, the pair percentage corresponding to the bucket (the value on the y-axis) does. Fig. 5 also shows an interesting characteristic that the distribution has the behaviour of an exponential function (the dotted curve) $p = C_1 e^{(C_2 \times d)}$ where C_1 and C_2 are constants and we experimentally found them to be: $C_1 = 1/N$ and $C_2 = 6.92$ where $N = 8{,}860$. In other words, the *GEOSO* model shows that the percentage of pairs increases *exponentially* as the social distance increases.

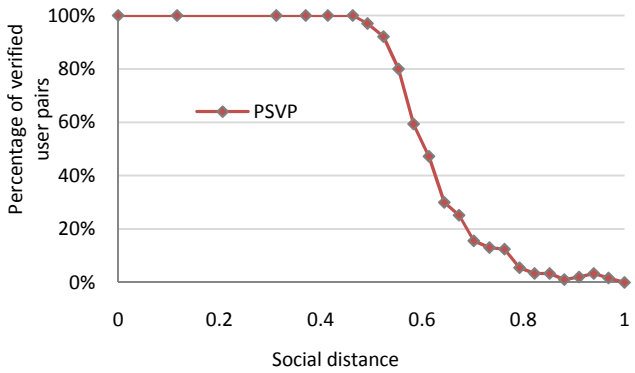

Fig. 6. PSVP vs. social distances – set of 2,000 people

Next, we verify the distances using the social information retrieved from IMDB and Wikipedia. The verified results are shown in Fig. 6. The x-axis shows the social distances and the y-axis represents the percentage of successfully verified pairs (PSVP) of each individual bucket for all 280k pairs. As Fig. 6 shows, buckets with distances less than 0.55 have PSVP above 80% (150 pairs), and buckets with distance less than 0.6 have PSVP above 59% (301 pairs). As the distance values increase, especially close to the value of 1, the percent of verified user pairs drops dramatically. This is due to the fact that when two persons are far away in social distances, there is no data from IMDB or Wikipedia showing that these two are not friends, family members or in other relationships, which on the other hand proves that our distance measure is consistent with the reality.

We also verify our results by manually checking the first 300 celebrity pairs with the smallest social distances. The user pairs with smallest distances are, for example, twins Close Sprouse and Dylan Sprouse, twins Ashley Olsen and Mary-Kate Olsen, and Ricky Gervais and Steven Merchant. These user pairs acted together in either many TV Series or movies.

In the next set of experiments, we use the dataset of 4,000 people and 1.1M co-occurrence events. The trends of the figures are similar to the previous set of experiments. Fig.7 shows the relationship between social distances and user pair percentage of 50k pairs corresponding to 4.5% of 1.1M pairs. Again, user pairs with

distances greater than 0.9 are dropped as they are considered socially far away. The x-axis shows social distances and the y-axis shows the percentage of pairs in buckets.

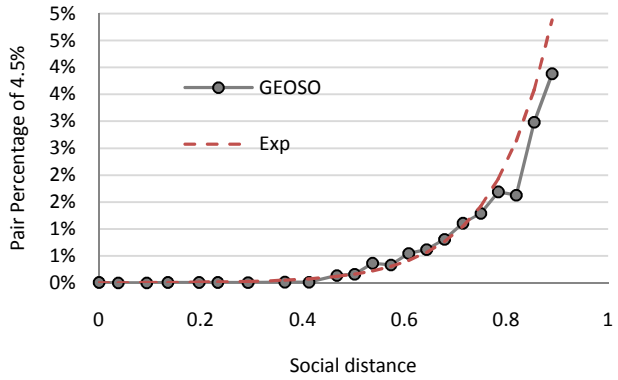

Fig. 7. Percentage of pairs vs. social distances – set of 4,000 people

As shown in Fig. 7, the graph also exhibits the behaviour of an exponential function $p = C_1 e^{(C_2 \times d)}$. The constants are $C_1 = 1/N$ and $C_2 = 8.76$ where $N = 8{,}860$. This behaviour holds the best for the buckets with distances less than 0.8. Beyond this point, the higher the distance, the more different the distribution is from its approximated exponential behaviour. This can be explained by the fact that when the size of the set increases, less famous people are added to the set and they acted in less movies/episodes than the more famous ones, hence they have less chance to act together with other people, which results in a sparser social graph and higher numbers of pairs falling into buckets of higher distances.

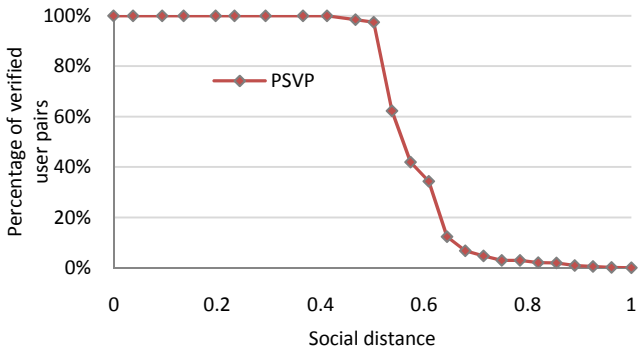

Fig. 8. PSVP vs. social distances – set of 4,000 people

Fig. 8 shows the realtionship between verified user pairs and distances. The x-axis shows the social distances, and the y-axis shows the PSVP values. When the social distance increases, the percentage of PSVP also decreases.

In the last set of experiments, we use a dataset that contains 10,000 people and 4.6M co-occurrence events. Fig. 9 shows the relationship between the percentage of pairs (the y-axes) and the social distance (the x-axis) for top 0.5% user pairs out of 4.6M pairs.

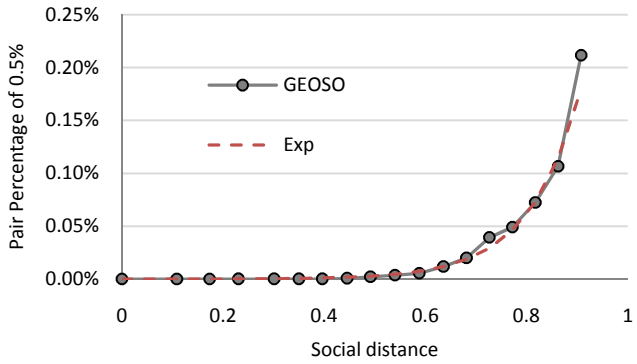

Fig. 9. Percentage of pairs vs. social distances – set of 10,000 people

Fig. 9 shows that the majority user pairs do not have close social connection due to the fact that in a large scale social network, most people are not directly socially connected.

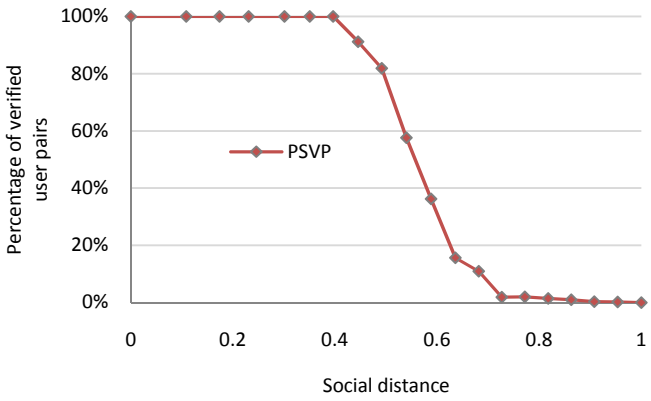

Fig. 10. PSVP vs. social distances – set of 10,000 people

In Fig. 10, we show the relationship between PSVP (y-axis) and social distance (x-axis). Similarly, it shows that user pairs with small distances are verified by the data available. Hence, user pairs that are considered socially close by our model are indeed close in reality.

5.3 Precision and Recall

In this section, we measure our results using the precision and recall model. For each value of the social distance d (the midpoint of each bucket), we calculate the precision and recall for the set of all pairs with social distance less than or equal to d.

$$Precision(d) = \frac{NSVP(d)}{NP(d)}, Recall = \frac{NSVP(d)}{\sum_d NSVP(d)} \quad (5)$$

The term $NSVP(d)$ represents the number of successfully verified pairs with distance no greater than d, and $NP(d)$ is the number of pairs with distance no greater than d. We present the precision and recall for all three datasets used in the previous sections. In Fig. 11, the x-axis shows the social distance and the y-axis shows the precision and recall measures.

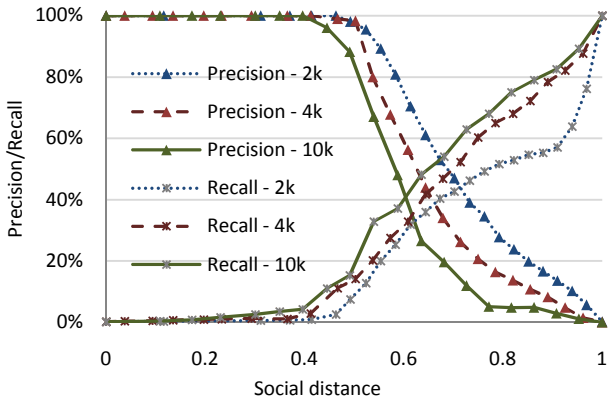

Fig. 11. Precision/Recall vs. Social Distance

As shown in Fig. 11, the precision is high as the distance values are small. However, the recall is low. This is due to two reasons. First, the number of pairs with shorter social distances is only a small fraction (3%-5%) of the total number of user pairs. Although all of them can be verified, they account for only a small percent of all user pairs. Second, user pairs who are close in reality are not reported close in our model. This is because our datasets consist of only co-acting data instead of real spatiotemporal co-occurrence data. Two persons who are father and son might never act together, but they are socially close. This generally is not the case in a real spatiotemporal dataset.

6 Conclusion and Future Work

In this paper, we focused on how to infer social connections among people based on their co-occurrences in space and time. We presented the *GEOSO* model which derives social connections between people based on spatiotemporal events in real

world. We also showed that our model captures the intuitive properties of social behaviors. Finally, our experiments demonstrated that the social distances computed by our model are consistent with the real social distances from the datasets.

There are a few future extensions for this work. First, we plan to extract more features from co-occurrence events, for example, the real distances between visits happened in the same cell and the overall time overlaps spent at same locations between two users. Then we can use these features to increase the precision of our social distance measure. Furthermore, once a social closeness is identified, we can also use the geospatial information and time to label the relationship. For example, if two persons go to only work-related places like an office building, a parking garage, and a nearby cafeteria during working hours, they are most likely colleagues. If two persons go to shopping malls, groceries and play sports together, it is more probable that they are friends or life partners.

Acknowledgments. This research has been funded in part by NSF grant CNS-0831505 (CyberTrust) and IS-1115153, the USC Integrated Media Systems center (IMSC), and unrestricted cash and equipment gift from Google, Microsoft and Qualcomm. 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 the National Science Foundation.

7 References

1. Crandall, D., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., Kleinberg, J.: Inferring social ties from geographic coincidences. In: Proc. National Academy of Sciences 107(52) 22436-22441(2010)
2. Shahabi, C.,Banaei-Kashani, F.: Efficient and anonymous web usage mining for web personalization. INFORMS Journal on Computing-Special Issue on Data Mining, Vol.15. No.2 (2003)
3. Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., Ma, W. Y.: Mining User Similarity Based on Location History. In: Proc. Of the 16th ACM SIGSPATIAL International Conference on Advances of GIS, New York, NY (2008)
4. Storch, H.,Zwiers, F.: Statistical Analysis in Climate Research. Cambridge University Pr. ISBN 0521012309(2001)
5. Diaconis, P., Mosteller,F.: Methods for Studying Coincidences. J Am Stat Assoc 84:853-861 (1989)
6. Backstrom, L., Dwok, C., Kleinberg, J.: Wherefore Art Throu R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography. In: Proc. of the 16th International World Wide Web Conference(2007)
7. Schaff, D. P., Waldhauser, F.: Waveform cross-correlation-based differential travel-time measurements at the northern California seismic network. Bull. Seism. Soc. Am. *96*, 38-49 (2006)

8. M. Rossi, T. M., Warner, I. M.: Pattern Recognition of Two-Dimensional Fluorescence Data Using Cross-Correlation Analysis. Applied Spectroscopy, Vol. 39, Issue 6, pp. 949-959(1985)
9. Yuan, S. T., Sun, J.: Ontology-based structured cosine similarity in document summarization: with applications to mobile audio-based knowledge management. In: IEEE Transaction on Systems, Man, and Cybernetics-Part B: Cybernetics, Vol. 35, No. 5 (2005)
10. Esteva, M., Bi, H.: Inferring Intra-organizational Collaboration from Best-matched Cosine Similarity Distributions in Text. In: Proc. of the 9th ACM/IEEE-CS joint conference on Digital libraries, JCDL (2009)
11. Zhang, D., Du, Y., Hu,L.: On Monitoring the top-k Unsafe Places. In: Proc. of 24th International Conference on Data Engineering (ICDE), Cancun, Mexico(2008)
12. Flicker, http://www.flickr.com/
13. LiveJournal, http://www.livejournal.com
14. IMDB, http://www.imdb.com
15. Kumar, B. V., Savvides, M., Xie, C.: Correlation pattern recognition for face recognition. In: Proc. Of the IEEE (2006)
16. Yuan, S. T., Sun, J.: Ontology-based Structured Cosine Similarity in Document Summarization: with Applications to Mobile Audio-Based Knowledge Management. IEEE Transaction on Systems, Man, and Cybernetics-Part B: Cybernetics, Vol. 35, No. 5 (2005)
17. Griffiths, T., Tenenbaum, J.: Randomness and Coincidences: Reconciling Intuition and Probability Theory. In: Proc. of the 23rd Annual Conference of the Cognitive Science Society pp 370-375 (2001)
18. Wikipedia, http://www.wikipedia.org
19. Pham, H., Hu, L., Shahabi, C.: Towards Integrating Real-World Spatiotemporal Data with Social Networks. In: Proc. Of the 19th ACM SIGSPATIAL International Conference on Advances in GIS, Poster Presentation, Chicago, Illinois (2011)
20. Yoon, H., Shahabi, C.: Accurate Discovery of Valid Convoys from Moving Object Trajectories. In: International Workshop on Spatial and Spatiotemporal Data Mining (SSTDM-09), Miami, Florida, USA (2009)
21. Lee, J. G., Han, J., Whang, K. Y.: Trajectory Clustering: a Partition-and-Group Framework. In: SIGMOD Conference pp 593-604 (2007)
22. Lee, J. G., Han, J., Li, X., Gonzalez, H.: TraClass: Trajectory Classification using Hierarchical Region-Based and Trajectory-Based Clustering. In: Proc. Of the VLDB Endowment, v.1 n.1 (2008)
23. Vieira, M. R., Bakalov, P., Tsotras, V. J.: On-line Discovery of Flock Patterns in Spatio-Temporal Data. GIS pp 286-295 (2009)
24. Roh, G. P., Roh, J. W., Hwang, S. W., Yi, B. K.: Supporting Pattern Matching Queries over Trajectories on Road Networks. TKDE (2010)