Towards Integrating Real-World Spatiotemporal Data with Social Networks

Huy Pham Ling Hu Cyrus Shahabi

Integrated Media Systems Center
University of Southern California (USC)
Los Angeles, California
{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 measures either do not scale or may overestimate the strength of social connections by giving too much weight to coincidences.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - *data mining*. H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval model*

Keywords

Spatiotemporal, geospatial, social networks, social relationship

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. 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 available. Those data in their simplest form captures the people *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 among people in their real lives [3, 4]. 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 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, taking into account both spatial and temporal factors. However, they do not consider the frequency of co-occurrences in space and time, and made a 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.

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 [5, 6] to measure the similarity between two time-series of two users. However, this approach would not scale well and would reflect a false notion of continuity of space, resulting in misrepresentation of the visit information in time intervals between two visits. Another tempting solution is to model a person's visit pattern as a vector where each dimension corresponds to a fixed location ID and the value capture the frequency of visits, and then use the cosine similarity [2, 7] to calculate the distance between two patterns represented by vectors. However, 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

ACM SIGSPATIAL GIS '11, November 1–4, 2011. Chicago, IL, USA

Social Model (*GEOSO*), towards integrating real-world spatiotemporal data with social-networks. We discuss various auxiliary representations such as *co-occurrence vector* and *master vector*, to enable an accurate distance computation.

The remainder of this paper is organized as follows. Section 2 formally defines the 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. Finally, we conclude the paper with future directions in Section 5.

2. PROBLEM DEFINITION

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). Two people, who visited multiple locations, or repeatedly visited the same location at the same time, are socially connected with higher probability. 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.

3. THE GEOSO MODEL

To better capture the relationship between spatiotemporal co-occurrences and social ties between 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, 8], 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 x Y rectangular cells. The size of the cells is application-dependent.

3.1.1 Visit vector

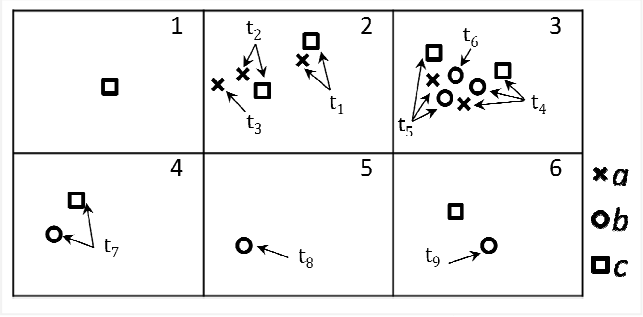

Figure 1. Visit history of user a, b and c.

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. 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.

For example, in Figure 1, the visit vectors of user *a* and user *b* are:

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

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

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. The value of each dimension records the number of times that the two users visited the same cell at *roughly the same time*. Note the length of the time overlap is application dependent and can be an input parameter to our model. Consider users *a* and *c* in Figure 1, *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, the 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*.

3.1.3 Master Vector

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. 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.

$$M = (m_1, m_2, \dots, m_k, \dots, m_N) \quad (2)$$

$$m_k = \max_{1 \le i < j \le U, 1 \le k \le N} c_{ik,jk}$$

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.

$$d_{ij} = \sqrt{\sum_k (c_{ik,jk} - m_k)^2} \quad (3)$$

$$s_{ij} = \frac{1}{d_{ij}}$$

Consider a simple example consisting of two cells and three users shown in Figure 2. 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 $M = (2,3)$.

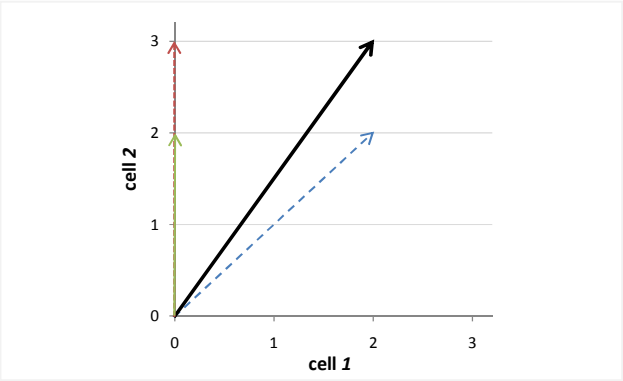

Figure 2. Vector view of GEOSO distance measurements.

Next, the PED distance between each user pair is computed as the distance from the master vector to the co-occurrence vector. The smaller the distance, the closer two users are socially.

4. PROPERTIES OF THE GEOSO MODEL

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, 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 assign the value *1* for that cell, and assign the value *0* otherwise. 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. 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:

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

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

$$= -a(m-1)^2 + am^2 = a(2m-1)$$

As m is greater than zero, we know $d_{ij}^2 > d_{pq}^2$. Hence d_{ij} is greater than d_{pq}. Consequently, user p and q are more socially connected than user i and j. Therefore, our model has the compatibility property.

4.2 Commitment

As stated in our second observation, if two users repeatedly visited the same place together, they are more likely socially close to each other. To show that our distance model is consistent with this observation, we need to take into account the number (frequency) of co-occurrences between two users which we left behind in the previous section. Then the second observation states that the more committed two users 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:

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

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

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

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

As the next step, we analyze the relationship between the two in the model and show which of the two properties is 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:

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

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.

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

In the equation above, m is a constant. The relationship between the variable x and variable y is plotted in Figure 3 (m is set to *20*).

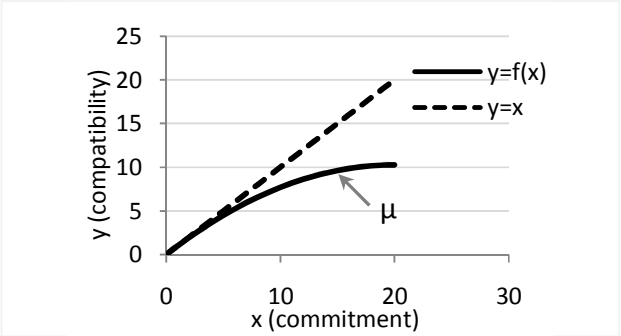

Figure 3. Commitment v.s. compatibility.

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 reality because multiple co-occurrences at a single location might just be an indicator of coincidences [9], such as students study in the same library and they are not friends of each other, while co-occurrences at multiple locations are seldom coincidences.

Second, it is shown in the Figure 3 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 coincidences 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. CONCLUSIONS 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. We leave the experiment for the future work, for which we plan to collect a large set of geospatial data that have information about the locations that people have been to, and the social connections among those people, which will be used to test the result of the model. We also plan to extract more features from co-occurrence events, such as 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.

6. 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] D. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher and J. Kleinberg, "Inferring social ties from geographic coincidences," Proc. National Academy of Sciences 107(52) 22436-22441, 28 December 2010.

[2] C. Shahabi, and F. Banaei-Kashani "Efficient and anonymous web usage mining for web personalization," INFORMS Journal on Computing-Special Issue on Data Mining, Vol.15. No.2, Spring 2003 .

[3] P. Diaconis, and F. Mosteller, Methods for Studying Coincidences, J Am Stat Assoc 84:853-861, 1989.

[4] L. Backstrom, C. Dwok, and J. Kleinberg, "Wherefore art throu R3579X? anonymized social networks, hidden patterns, and structural steganography," Proc. of the 16th International World Wide Web Conference, 2007.

[5] H. Storch, and F. Zwiers, Statistical Analysis in Climate Research, Cambridge University Pr. ISBN 0521012309, 2001.

[6] B.V. Kumar, M. Savvides, and C. Xie, "Correlation pattern regconition for face recognition," Proc. Of the IEEE, Nov. 2006.

[7] S.T. Yuan, and J. Sun, "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, October 2005.

[8] D. Zhang, Y. Du, and L. Hu, "On Monitoring the top-k Unsafe Places", Proc. of 24th International Conference on Data Engineering (ICDE), Cancun, Mexico, 2008.

[9] T. Griffiths, and J. Tenenbaum, "Randomness and coincidences: reconciling intuition and probability theory," Proc. of the 23rd Annual Conference of the Cognitive Science Society pp 370-375.