

When and Where: Predicting Human Movements Based on Social Spatial-Temporal Events

Ning Yang*

Xiangnan Kong[†]

Fengjiao Wang[‡]

Philip S. Yu[§]

Abstract

Predicting both the time and the location of human movements is valuable but challenging for a variety of applications. To address this problem, we propose an approach considering both the periodicity and the sociality of human movements. We first define a new concept, Social Spatial-Temporal Event (SSTE), to represent social interactions among people. For the time prediction, we characterise the temporal dynamics of SSTEs with an ARMA (AutoRegressive Moving Average) model. To dynamically capture the SSTE kinetics, we propose a Kalman Filter based learning algorithm to learn and incrementally update the ARMA model as a new observation becomes available. For the location prediction, we propose a ranking model the periodicity and the sociality of human movements are simultaneously taken into consideration for improving the prediction accuracy. Extensive experiments conducted on real data sets validate our proposed approach.

1 Introduction

Recently, predicting human movements has been attracting considerable interest because of its great value in a variety of applications such as location-based services, monitoring of epidemic[23, 8, 21]. Existing researches share a common idea to model human movements based on sequential patterns discovered from historical trajectories[14, 16, 10, 17, 6, 2], and have two major shortcomings. Firstly, existing approaches mainly focus on predicting the next locations only, but neglect the occurrence time. In many real-world applications, it is usually essential to know both the next location where a given person will appear, and the time when he/she will appear at that location. For example, in homeland security, it is very important to estimate both when and where a suspect will appear. In epidemic monitoring,

accurate prediction of both when and where people will be infected by a kind of deadly virus is vital for the prevention of disease outbreak. Secondly, existing prediction models based on sequential patterns are not easy to understand while assuming the activities of different people are independent from each other.

In fact, more and more research evidences show that most of human movements are strongly periodic, and can be understood in the context of social relationships [8, 21, 11, 19, 4, 9, 23]. For example, when people go out, they are more likely to party with friends, or meet with colleagues than going to a random place, and these social interactions often take place regularly. These research evidences and observations suggest that the prediction of human movements can be largely reduced to the prediction of social interactions.

Motivated by the above applications and observations, in this paper we aim at the problem of predicting both the time and the location of a given person's next movement, by considering his/her social relationship information. For this purpose, there are two main issues we have to address:

- (1) Social relationships are generally represented as graphs, while the time and the location of human movements are usually modeled as time series or sequences. So, how to utilize social information, the graph data, to facilitate the prediction of time and location, the sequence data?
- (2) Intuitively, it is hard to estimate the time and the location at the same time, so, which prediction should be made first, and how does one help the other?

For the first issue, we propose a new concept, *Social Spatial-Temporal Event (SSTE)*, to model social interactions between a given person and his/her friends. Essentially, an SSTE of a given person is a subgraph where nodes represent the given person and a subset of his/her friends. We associate each subgraph with two spatial-temporal attributes, i.e., the time when and the location where a social interaction occurs. SSTE integrates the social information and the spatial-temporal information of human movements, so the regularity of SSTEs of a given person can reveal the dynamics of so-

*College of Computer Science, Sichuan University, China.
yangning@scu.edu.cn

[†]Department of Computer Science, University of Illinois at Chicago, USA. xkong4@uic.edu

[‡]Department of Computer Science, University of Illinois at Chicago, USA. fwang27@uic.edu

[§]Department of Computer Science, University of Illinois at Chicago, USA. psyu@uic.edu

cial interactions of that person, which makes it reasonable and feasible to formulate the prediction of human movements as the time and location predictions of the next SSTE of a given person.

For the second issue, our idea is in light of the plenty of researches that indicate the human movement exhibits a strong periodicity [8, 21] and a strong sociality [19]. The periodicity implies that the location depends on its occurrence time. For example, the places where people meet on weekday are reasonably different than those on weekend. The sociality implies that the location of an SSTE of a given person is also influenced by the preference of his/her friends. That is to say the place where an event occurs will be influenced by not only the friends or the participants, but also the time of occurrence. Hence making time prediction first will be better able to exploit these properties. We thus make time prediction first, then use the predicted time combined with the social information as the input of the location prediction.

Time Prediction The goal of the time prediction of an SSTE is to estimate how long it will take before the next SSTE involving a given person occurs. We characterize the temporal dynamics of SSTEs of a given person with an ARMA (AutoRegressive Moving Average) model [5]. The time prediction based on the ARMA model only depends on the mean value and the autocorrelation function of the time interval series. This is extremely important for our goal because of the difficulty of estimating all of the joint distributions on massive data.

As user behavior can change from time to time, an ARMA model whose parameters keep constant will not be able to fully capture the behavior dynamics. Here we propose a learning algorithm which employs Kalman Filter [22] to learn and incrementally update the parameters of the prediction model as a new observation becomes available. Our learning algorithm requires only the latest observation to update the ARMA model, which is in sharp contrast with the traditional methods like MLE (Maximum Likelihood Estimation) that recalculate the parameters based on the whole data every time when a model update is required.

Location Prediction By combining the periodicity and sociality of human movements, we propose a ranking model by which we rank all the candidate locations and output the locations with higher ranking scores as the prediction result.

In summary, the contributions of this paper are as follows:

- (1) We reduce the prediction of human movements to the prediction of social interactions causing those movements. For this purpose, we propose a new

concept, Social Spatial-Temporal Event, to characterize the periodicity and sociality of human movements.

- (2) For the prediction of the occurrence time of an SSTE, we propose an ARMA based prediction model and a Kalman Filter based learning algorithm which is capable of incrementally updating the prediction model as a new observation becomes available.
- (3) We propose a ranking model for the prediction of an SSTE location, where the periodicity and the sociality of human movements are simultaneously taken into consideration for improving the prediction accuracy.
- (4) The performance of our proposed methods are validated via the extensive experiments conducted on real data sets.

The rest of this paper is organized as follows. The preliminary concepts and terminologies are introduced in Section 2. The concept of SSTE and its detection are described in Section 3. The time prediction model and algorithm are presented in Section 4. The ranking model and the location predicting algorithm are worked out in Section 5. We analyze the experimental results in Section 6. At last, we briefly review the related work in Section 7 and conclude in Section 8.

2 Preliminaries

In this section we briefly introduce the basic concepts of ARMA model and Kalman Filter.

2.1 AutoRegressive-Moving-Average Model

AutoRegressive-Moving-Average (ARMA) model is a powerful tool to analyze a stationary stochastic process. Given a time series of historical data $x^t = (x_1, x_2, \dots, x_t)$, where $x_i (1 \leq i \leq t)$ is the data vector at time i , the ARMA model can predict the future value \hat{x}_{t+1} . An ARMA model consists of two parts, an autoregressive (AR) part and a moving average (MA) part. The AR model of order p , AR(p), is given by the following equation:

$$(2.1) \quad \Phi(B)(x_t) = 0,$$

where

$$\Phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p,$$

and B is the backward shift operator, $B^j x_t = x_{t-j}$. The MA model of order q , MA(q), is given by the following equation:

$$(2.2) \quad \Theta(B)\varepsilon_t = 0,$$

where

$$\Theta(z) = 1 - \theta_1 z - \dots - \theta_q z^q,$$

and ε_t is a white noise with zero mean and variance σ^2 . Then by combining the equation (2.1) and (2.2), the ARMA model of order (p, q) , can written as:

$$(2.3) \quad \Phi(B)(x_t) = \Theta(B)\varepsilon_t.$$

2.2 Kalman Filter Kalman filter is a set of mathematical equations that provides an efficient computational (recursive) means to estimate the state of a linear dynamic process from a series of noisy measurements, in a way that minimizes the mean of the squared error. Kalman Filter assumes the true state at time k evolves from the previous state at $(k - 1)$, i.e.,

$$(2.4) \quad x_k = Ax_{k-1} + Bu_{k-1} + w_{k-1},$$

where u_{k-1} and w_{k-1} are the control input and the process noise respectively, A is the state transition matrix and B is the control input matrix. A might change over time, but here we assume it is constant. Since the control input is optional, so for simplicity, we assume $u_k = 0$ for any k . The process noise w_k is assumed to follow a normal distribution with zero mean and covariance Q_k ,

$$(2.5) \quad p(w_k) \sim N(0, Q_k).$$

At time k , an observation (or measurement) z_k of the true state x_k is

$$(2.6) \quad z_k = Hx_k + v_k,$$

where H is the measurement matrix and v_k is the observation noise which is also assumed to follow a normal distribution with zero mean and covariance R_k ,

$$(2.7) \quad p(v_k) \sim N(0, R_k).$$

Our objective is to predict $\hat{x}_{k|k}$, the posteriori state estimate at time k given observations up to k . This computation can be divided into two alternate phases, "Predicting" and "Updating". In the predicting phase, the priori state estimate $\hat{x}_{k|k-1}$ is produced from the posteriori state estimate at previous time step $\hat{x}_{k-1|k-1}$ [22], i.e.,

$$(2.8) \quad \hat{x}_{k|k-1} = A\hat{x}_{k-1|k-1} + Bu_{k-1},$$

and

$$(2.9) \quad P_{k|k-1} = AP_{k-1|k-1}A^T + Q_k,$$

where $P_{k|k-1}$ is the covariance matrix of the priori estimate.

In the updating phase, the current posteriori state estimate is generated by combining the current priori estimate with the current observation information.

Kalman Filter has a set of formulas well-established for the update phase, respectively shown as follows [22].

Innovation (or measurement residual):

$$(2.10) \quad y_k = z_k - H\hat{x}_{k|k-1}.$$

Innovation covariance:

$$(2.11) \quad S_k = HP_{k-1}H^T + R_k.$$

Optimal Kalman gain:

$$(2.12) \quad K_k = P_{k|k-1}H^T S_k^{-1}.$$

Posteriori state estimate:

$$(2.13) \quad \hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k y_k.$$

Posteriori estimate covariance:

$$(2.14) \quad P_{k|k} = (I - K_k H)P_{k|k-1}.$$

3 Social Spatial-Temporal Event

What LBSNs provide us contain a friendship graph and whole sequences of recorded checkins. The friendship graph is denoted by $G(V, E)$ where $G.V$ is the collection of persons and $G.E$ is the set of edges. There exists an edge between two persons if they are friends. The checkin sequence is denoted by $R = (r_1, r_2, \dots, r_n)$ where r_i is the i th checkin in order of time. The concept of checkin and Social Spatial-Temporal Event are defined as follows:

DEFINITION 3.1. Checkin a checkin record r is a 3-tuple (v, τ, σ) , where $r.v \in G.V$ is the person who checks in, $r.\tau$ is the checkin time, $r.\sigma$ is the coordinates of the checkin location.

DEFINITION 3.2. Social Spatial-Temporal Event (SSTE) an SSTE, denoted by a , is a 3-tuple (G, τ, σ) , where $a.G$ is a subgraph of the friendship graph G , $a.\tau$ is the occurrence time of a , and $a.\sigma$ is the coordinates of the location where a occurs.

There exist a lot of time-evolving clustering algorithms that can serve the detection of SSTEs from a checkin sequence. In this paper, we implement the SSTE detection by the density based temporal clustering algorithm proposed by Yixin Chen et al. [7], because its time-evolving weight schema is suitable for the requirement to adaptively determine the temporal boundary between two successive SSTEs.

4 Time Prediction

Now we turn our attention to the time prediction of SSTE, which can be formalized as follow:

Time Prediction With the known SSTE sequence involving person u , denoted as

$$A_u^n = (a_{u,1}, a_{u,2}, \dots, a_{u,n}), \forall a_{u,i}, 1 \leq i \leq n, u \in a_{u,i}.V$$

where $a_{u,i}$ is the i th SSTE of u , we want to predict the occurrence time of the next SSTE of u , i.e., $a_{u,n+1}.\tau$.

The time interval series x_u^t can be derived from A_u^n , i.e. $x_u^t = (x_{u,1}, x_{u,2}, \dots, x_{u,t})$, where $t = n - 1, x_{u,i} = a_{u,i+1}.\tau - a_{u,i}.\tau, 1 \leq i \leq t$. Now the original problem turns out to be the prediction of $x_{u,t+1}$ with known x_u^t .

4.1 Model Selection For a given person u , we can model x_u^t by an ARMA(p, q) model given by equation (2.3):

$$(4.15) \quad \Phi(B)(x_{u,t}) = \Theta(B)\varepsilon_{u,t},$$

where

$$\begin{aligned} \Phi(z) &= 1 - \phi_{u,1}z - \cdots - \phi_{u,p}z^p, \\ \Theta(z) &= 1 - \theta_{u,1}z - \cdots - \theta_{u,q}z^q, \end{aligned}$$

and B is the backward shift operator, $B^j x_{u,t} = x_{u,t-j}$, and $\varepsilon_{u,t}$ is a white noise with zero mean and variance σ_u^2 .

The task of model selection is to determine the orders p and q , which can be fulfilled by the following canonical procedure defined in time series analysis[5]:

- (1) If the ACF (Auto-Covariance Function) of the x_u^t cuts off after lag q and the PACF (Partial ACF) tails off, x_u^t is modeled as ARMA(0, q);
- (2) If the PACF cuts off after lag p and the ACF tails off, x_u^t is modeled as ARMA(p , 0);
- (3) If both the ACF and the PACF tail off, x_u^t is modeled as ARMA(p, q) where p and q are determined by the AIC (Akaike's Information Criterion) [5].

Note that x_u^t should be substituted with its first or even higher order difference series if x_u^t is not stationary, since the above procedure can apply only to a stationary process.

4.2 Learning the Model Once the model's orders are determined, the parameters, $\phi_{u,1}, \dots, \phi_{u,p}, \theta_{u,1}, \dots, \theta_{u,q}$ have to be learned. To make Kalman Filter applicable to our situation, the challenge here is to establish our own *measurement equation* and *state equation* [22].

Measurement Equation By equation (2.1), we have

$$(4.16) \quad x_{u,t} = H_{u,t}^T \Phi_{u,t} + \Theta_u^T \varepsilon_{u,t},$$

where

$$\begin{aligned} H_{u,t} &= [x_{u,t-1}, x_{u,t-2}, \dots, x_{u,t-p}]^T, \\ \Phi_{u,t} &= [\phi_{u,1}, \dots, \phi_{u,p}]^T, \\ \Theta_u &= [1, -\theta_{u,1}, \dots, -\theta_{u,q}]^T, \\ \varepsilon_{u,t} &= [\varepsilon_{u,t}, \varepsilon_{u,t-1}, \dots, \varepsilon_{u,t-q}]^T. \end{aligned}$$

In light of the theory of Bayesian statistics, our idea here is to regard the parameter vector $\Phi_{u,t}$ as the random vector representing the state of the underlying system, and $x_{u,t}$ as the observable variable (i.e., $\Phi_{u,t}$ plays the role of x_k in equation (2.4) and $x_{u,t}$ plays the role of z_k in equation (2.6)). We hence let equation (4.16) be the measurement equation of Kalman Filter, where time-varying matrix $H_{u,t}$ is the measurement matrix, and $\Theta_u^T \varepsilon_{u,t}$ is the measurement noise with zero mean and the variance

$$r_u = (1 + \sum_{i=1}^q \theta_{u,i}^2) \sigma_u^2.$$

State Equation The state equation is a bit less obvious. Considering equation (4.15) is an ARMA model, we deliberately establish the state equation as

$$(4.17) \quad \Phi_{u,t} = \Phi_{u,t-1} + W_{u,t},$$

where the $p \times 1$ matrix $W_{u,t}$ is the process noise which is assumed to be independent, white, and with zero mean and the covariance Q_u , a $p \times p$ matrix.

Let $\hat{\Phi}_{u,t}$ be the estimate of $\Phi_{u,t}$, $e_{u,t}$ be the estimate error,

$$e_{u,t} = \Phi_{u,t} - \hat{\Phi}_{u,t}.$$

Let $P_{u,t}$ be the covariance matrix of the estimate error, i.e.,

$$P_{u,t} = E[e_{u,t} e_{u,t}^T].$$

By combining the time update procedure and the measurement update procedure in standard Kalman Filter algorithm [22], we have the following iterative equations for the parameter update when a new observation $x_{u,t}$ becomes available:

(4.18)

$$K_{u,t+1} = (P_{u,t} + Q_u) H_{u,t} [H_{u,t}^T (P_{u,t} + Q_u) H_{u,t} + r_u]^{-1}$$

$$(4.19) \quad P_{u,t+1} = (I - K_{u,t+1} H_{u,t}^T) (P_{u,t} + Q_u)$$

$$(4.20) \quad \hat{\Phi}_{u,t+1} = \hat{\Phi}_{u,t} + K_{u,t+1} (x_{u,t} - H_{u,t}^T \hat{\Phi}_{u,t}),$$

where I is the unit matrix and $K_{u,t+1}$ is the Kalman gain, a $p \times 1$ matrix.

Initialization We use traditional MLE once to initialize $\hat{\Phi}_{u,0}$, Θ_u , and σ_u^2 . As for any $P_{u,0} \neq 0$ and

for any Q_u the estimate given by equation (4.20) would eventually converge [22], we simply presume $P_{u,0} = 1_{p \times p}$ and $Q_u = 0_{p \times p}$.

As the result of this subsection, the parameter learning algorithm is given in Algorithm 1.

Algorithm 1 *LearnByKF*($\hat{\Phi}_{u,t}, x_{u,t}, H_{u,t}, P_{u,t}$)

Require:

- $\hat{\Phi}_{u,t}$: the estimator of the last time;
- $x_{u,t}$: the new observation;
- $H_{u,t}$: the historical observations;
- $P_{u,t}$: the covariance of estimate error at time t ;

Ensure: $\hat{\Phi}_{u,t+1}$: the new estimate of parameters;
 1: Update the Kalman gain $K_{u,t+1}$ by equation (4.18);
 2: Update the covariance $P_{u,t+1}$ by equation (4.19);
 3: Generate the new estimate $\hat{\Phi}_{u,t+1}$ by equation (4.20);

4.3 Making Time Prediction In this paper, we define the best estimate of $x_{u,t+1}$, $\hat{x}_{u,t+1}$, to be the predictor with minimum mean square error, i.e., given the history X_u^t ,

$$\hat{x}_{u,t+1} = E[X_{u,t+1}|X_u^t].$$

where X_u^t represents the random sequence corresponding to x_u^t , i.e., $X_u^t = (X_{u,1}, X_{u,2}, \dots, X_{u,t})$. Concretely, we can compute $\hat{x}_{u,t+1}$ in terms of the following theorem.

THEOREM 4.1. *The best estimate of $x_{u,t+1}$ is*

$$(4.21) \quad \hat{x}_{u,t+1} = \hat{\Phi}_{u,t+1}^T H_{u,t+1} - \sum_{i=1}^q \theta_{u,i} \varepsilon_{u,t-i+1}.$$

Proof. By equation (4.16),

$$X_{u,t+1} = \sum_{i=1}^p \hat{\phi}_{u,i} X_{u,t-i+1} + \varepsilon_{u,t+1} - \sum_{i=1}^q \theta_{u,i} \varepsilon_{u,t-i+1}.$$

So,

$$\begin{aligned} \hat{x}_{u,t+1} &= \sum_{i=1}^p \hat{\phi}_{u,i} E[X_{u,t-i+1}|X_u^t] + E[\varepsilon_{u,t+1}|X_u^t] \\ &\quad - \sum_{i=1}^q \theta_{u,i} E[\varepsilon_{u,t-i+1}|X_u^t]. \end{aligned}$$

Since $\varepsilon_{u,t}$ and $\varepsilon_{u,t-i+1}$ ($1 \leq i \leq q$) are independent of X_u^t , we have

$$E[X_{u,t-i+1}|X_u^t] = x_{u,t-i+1},$$

$$\begin{aligned} E[\varepsilon_{u,t+1}|X_u^t] &= 0, \\ E[\varepsilon_{u,t-i+1}|X_u^t] &= \varepsilon_{t-i+1}. \end{aligned}$$

Hence substituting them into the above equation gives equation (4.21).

Since $\hat{x}_{u,t+1}$ is just the estimate of the time-gap between $a_{u,t}$ and $a_{u,t+1}$, the occurrence time of $a_{u,t+1}$ will be further estimated by

$$(4.22) \quad a_{u,t+1} \cdot \hat{\tau} = a_{u,t} \cdot \tau + \hat{x}_{u,t+1}.$$

Algorithm 2 gives the complete procedure of time prediction.

Algorithm 2 *AlgKF*($\hat{\Phi}_{u,t}, x_u^t, H_{u,t}, P_{ut}$)

Require:

- $\hat{\Phi}_{u,t}$: the last estimate of the parameters;
- x_u^t : the historical observations;
- $H_{u,t}$: the measurement matrix;
- $P_{u,t}$: the covariance of estimate error at time t ;

Ensure: $a_{u,t+1} \cdot \hat{\tau}$: the estimate of the occurrence time of $a_{u,t+1}$;

- 1: $\hat{\Phi}_{u,t+1} = \text{LearnByKF}(\hat{\Phi}_{u,t}, x_{u,t}, H_{u,t}, P_{u,t})$;
 - 2: **for** $i = 1$ to q **do**
 - 3: $\varepsilon_{u,t-i+1} = x_{u,t-i+1} - \hat{x}_{u,t-i+1}$;
 - 4: **end for**
 - 5: Compute $\hat{x}_{u,t+1}$ in terms of equation (4.21);
 - 6: Compute $a_{u,t+1} \cdot \hat{\tau}$ in terms of equation (4.22);
-

Note that if x_u^t is substituted with its difference series in equation (4.15), we have to further correct $\hat{x}_{u,t+1}$ with the differences just before line 6.

5 Location Prediction

The problem of the location prediction can be formalized as follow:

Location Prediction Given A_u^n , u 's friend set $F(u)$, and the predicted occurrence time $a_{u,n+1} \cdot \hat{\tau}$, predict the location $a_{u,n+1} \cdot \sigma$.

In order to predict the location at the future time point $a_{u,n+1} \cdot \hat{\tau}$ predicted by the Kalman Filter based algorithm described in Section 3, we explore the concepts of the periodic behavior and the influence of friends.

First, we consider the periodic behavior. We use one week as the period and one hour as the granularity and convert $a_{u,n+1} \cdot \hat{\tau}$ to the hour of a weekday. For example, if the predicted time is noon next Friday, we will use the noon time locations of previous Fridays to help make the prediction. Let L_u and T_u be the random variables respectively representing the location and the time of an SSTE involving u . We define the temporal

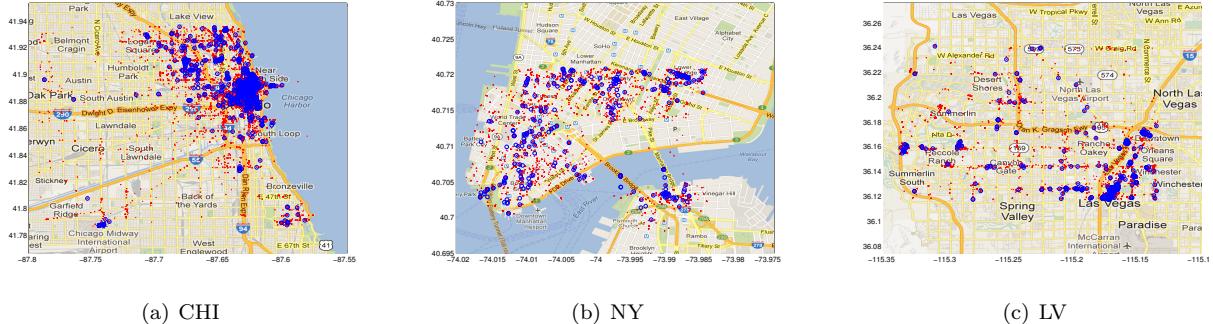


Figure 1: Geographic distribution of checkins and SSTEs.

score of a candidate location λ as:

$$(5.23) \quad g_{temporal}(\lambda) = P(L_u = \lambda | T_u = \hat{\tau}) = \frac{P(T_u = \hat{\tau} | L_u = \lambda)P(L_u = \lambda)}{P(T_u = \hat{\tau})},$$

where $\hat{\tau} = a_{u,n+1} \cdot \hat{\tau}$, and the probabilities can be respectively estimated by the following equations:

$$P(T_u = \hat{\tau}) = \frac{|\{a : a \in A^n \wedge a.\tau = \hat{\tau}\}|}{n},$$

$$P(L_u = \lambda) = \frac{|\{a : a \in A^n \wedge a.\sigma = \lambda\}|}{n},$$

$$P(T_u = \hat{\tau} | L_u = \lambda) = \frac{|\{a : a \in A^n \wedge a.\tau = \hat{\tau} \wedge a.\sigma = \lambda\}|}{|\{a : a \in A^n \wedge a.\sigma = \lambda\}|}.$$

Next, we explore the influence of friends, which means the location of $a_{u,n+1}$ is also dependent on the locations where any of u 's friends previously checked in before $a_{u,n+1} \cdot \hat{\tau}$. In addition, it is reasonable that more recent checkins are more influential. Hence we define the social score of a candidate location λ as

$$(5.24) \quad g_{social}(\lambda) = \frac{W_F}{W},$$

where W_F is the total weight of u 's friends who previously checked in λ before $\hat{\tau}$,

$$W_F = \sum_{r \in R \wedge r.v \in F(u) \wedge r.\sigma = \lambda} r.w(\hat{\tau}),$$

and W is the total weight of all of u 's friends,

$$W = \sum_{r \in R \wedge r.v \in F(u)} r.w(\hat{\tau}).$$

By combining equations (4.9) and (4.10), the overall ranking score of a candidate location is defined as

$$(5.25) \quad g(\lambda) = \xi g_{temporal}(\lambda) + (1 - \xi) g_{social}(\lambda),$$

where $\xi \in [0, 1]$ is the weight reflecting how important the periodicity is. In practice, we choose all the locations where u and any one of u 's friends previously checked in before $a_{u,n+1} \cdot \hat{\tau}$ as the candidate location collection.

At last, note that it is hard and pointless to predict the accurate coordinates of an SSTE location. So as a preprocessing step before making location prediction, we discretize the coordinates of locations to regions by Voronoi algorithm [3], where the regions are defined by the locations of bus stops. Then we practically predict the region instead of the coordinates.

6 Experimental Evaluation

The objectives of experiments are to verify the validity of the SSTE detection algorithm and respectively evaluate the prediction models for SSTE time and location. All the experiments are conducted on a PC with Intel Core i7 CPU 2.0G HZ, 8 GB main memory. All the algorithms are implemented in MATLAB 2011b.

Data Sets We use three real checkin data sets and the corresponding friendship graphs offered by [1], whose checkin areas are around Chicago (CHI), New York (NY) and Las Vegas (LV) respectively. CHI, NY and LV respectively contain 38,830, 2296, and 17,042 social checkins in one year. The data distributions are shown in Figure 1 (a), (b) and (c), where a red dot represents a checkin and a blue circle an SSTE.

6.1 Test of Time Prediction We compare AlgKF with AlgMLE, an algorithm we develop for the test. AlgMLE also makes prediction in terms of the equation (4.22) but the parameters keep constant all time once they have been learned by using MLE. AlgKF and AlgMLE run over CHI, NY and LV with different length proportions of training sequences ranged from 20% to 90%, respectively. We measure the accuracy by the average MSE (Mean Squared Error) on each training

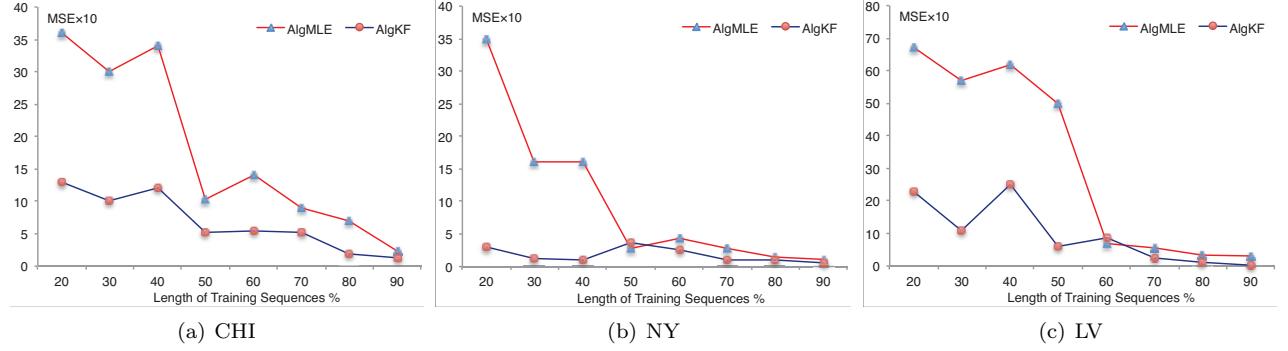


Figure 2: Error of time prediction.

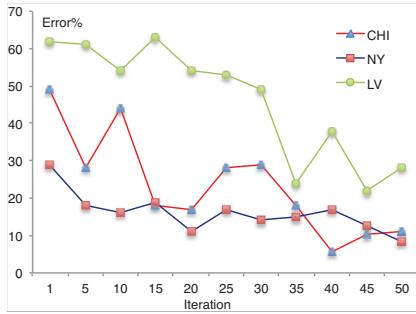


Figure 3: Converging speed of error.

data set. The results are shown in Figure 2, from which we can have the following observations and analyses:

- (1) MSE decreases with the increased length proportion of training sequences, whichever algorithm runs. This corresponds with our expectation that more historical data are used, more accurate predictions can be made.
- (2) The MSE incurred by AlgKF is generally lower than that incurred by AlgMLE at different length proportions of training sequences. This demonstrates that AlgKF outperforms the algorithms that keep the prediction model parameters constant because AlgKF is able to dynamically update the prediction model by Kalman Filter.
- (3) The MSE difference between AlgKF and AlgMLE is more significant at less length proportions of training sequences. The underlying reason is that although the initial accuracy is low due to the small training data, AlgKF is able to dynamically learn from the errors in virtue of Kalman Filter, while AlgMLE is unable. So AlgKF can quickly and continuously improve the prediction accuracy as more and more observations come into sight, which results in that the final MSE would not be too large.

Figure 3 further illustrates the error convergence speed on three particular SSTE sequences chosen from CHI, NY and LV respectively, where the proportion of training data is 20%. As we can see, the error drops sharply after about 30 iterations compared with the error after first iteration.

6.2 Test of Location Prediction We use *Accuracy@TopN*, a metric proposed by [18], to assess the performance of the location prediction model (equation (12)). In terms of *Accuracy@TopN*, an SSTE location is predicted successfully if it is ranked in the Top- N list.

Figure 4 shows the average accuracy over the three data sets, at different Top- N list sizes and at different values of the weight ξ , where the length proportion of training sequences is set to 80%. At first, we can observe that the accuracy is increasing with increased N , no matter what value of ξ is, which logically corresponds with our expectation. Figure 4 also shows different values of ξ have different impacts on the accuracy. When $\xi = 0$, i.e., the location of the next SSTE of a person u is predicted completely by the preference of u 's friends, the accuracy is totally unacceptable, since even *Accuracy@Top50* is only around 10%. With increased ξ , i.e., more consideration is given to the periodicity, the accuracy increases, until $\xi = 0.8$. When $\xi = 1$, i.e., the location prediction is made completely in terms of the periodicity, the accuracy is lower than the accuracy at $\xi = 0.8$. So, we can draw a conclusion that the location of an SSTE depends on both the periodicity and the sociality, but more on periodicity, which verifies the assumption we make in Section 1.

At last, we investigate how the accuracy of location prediction changes with the length of training SSTE sequences. We show the *Accuracy@Top5* and the *Accuracy@Top20* at different lengths of training sequences in Figure 5(a) and (b) respectively. As we can observe from Figure 5, the accuracy generally in-

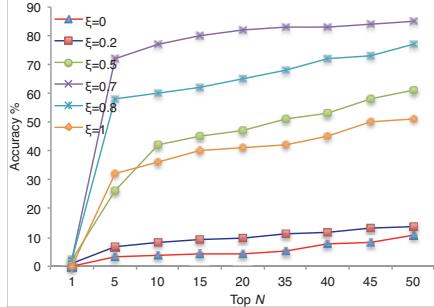


Figure 4: *Accuracy@TopN* on 80% training data.

creases with the increased length proportion of training sequences, which implies that longer the training sequence, more accurate the time prediction, and consequently more accurate the location prediction.

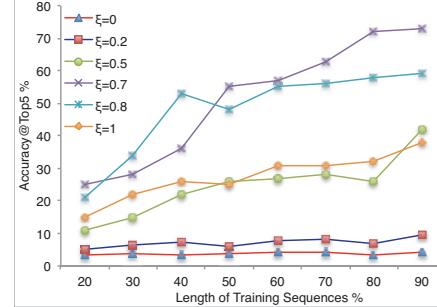
7 Related Work

There are three domains related to our work, Spatial-Temporal Event Mining, Time Prediction, and Location Prediction.

Spatial-Temporal Event Mining Hady W. Lauw et al. [13] propose a spatial-temporal event model to discover social associations among individuals based on spatial-temporal co-occurrences. The concept of the spatial-temporal event proposed by [13] is different from ours since it allows an individual to participate more than one event at the same time. Xingjie Liu et al. [15] identify a new type of social network, event-based social network (EBSN) and employ an extended Fiedler method for the community detection.

Time Prediction To our best knowledge, few of the previous studies focus on the prediction of the SSTE time. In [20], Yizhou Sun et al. propose a generalized linear model based model to predict the first building time of the link between two entities who have never been linked before, rather than how long the time is before the existing entities will be linked again in the future, while the latter is exactly the objective for the time prediction of our work. Zan Huang el al. [12] proposes a time series model to evaluate the frequency of repeated links, which focuses not on the prediction of the time when a link would occur, but on the prediction of whether a link would occur.

Location Prediction Location prediction has been received extensive attention in the field of trajectory analysis. Most of them are based on sequential pattern [11, 14, 16, 10], where the time just play the role of a time-stamp. Eunjoon Cho et al. [8] and Dashun Wang et al. [21] find that movements of humans are periodic and correlated with their social network. Eunjoon



(a) *Accuracy@Top5*

(b) *Accuracy@Top20*

Figure 5: Accuracy at different lengths of training sequences.

Cho et al. further reveal that human movements can be explained by periodic behavior more than by social relationships, and propose a periodic and social mobility model for predicting mobility of individuals, which focuses on predicting the location of a single checkin, rather than the location of a social interaction with at least two participants. Meanwhile it is based on the time given by user, not the time predicted in advance.

Note that our work is different with the traditional event prediction. The traditional event prediction focuses on the problem of whether an event will occur at a given time point, while our work focuses on both the occurrence time and the occurrence location.

8 Conclusions

In this paper, we reduce the problem of human movement prediction to the prediction of SSTEs. For the time prediction, we propose an ARMA based prediction model and a Kalman Filter based learning algorithm. Our proposed time prediction algorithm requires only the latest data and can hence incrementally update the prediction model as a new observation becomes available. For the location prediction, we propose a ranking model to predict the location of the SSTE at the predicted occurrence time, where the periodicity and the sociality of SSTEs are simultaneously taken into con-

sideration for improving the prediction accuracy. The experimental results demonstrate that our proposed algorithms can predict the time and location of an SSTE in an acceptable accuracy.

Acknowledgments

This work is supported by the National Science Foundation of China under Grant Nos. 61173099, 61103043 and the Doctoral Fund of Ministry of Education of China under Grant No. 20110181120062. This work is also supported in part by NSF through grants CNS-1115234, DBI-0960443, and OISE-1129076, and US Department of Army through grant W911NF-12-1-0066.

References

- [1] <http://snap.stanford.edu/data/loc-gowalla.html>.
- [2] O. Abul, F. Bonchi, and F. Giannotti. Hiding sequential and spatiotemporal patterns. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(12):1709–1723, 2010.
- [3] F. Aurenhammer. Voronoi diagrams – a survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23(3):345–405, 1991.
- [4] R. M. Bell, Y. Koren, and C. Volinsky. Modeling relationships at multiple scales to improve accuracy of large recommender systems. In *KDD’07*, pages 95–104, 2007.
- [5] P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting*. Springer, 2nd edition, 2002.
- [6] H. Cao, N. Mamoulis, and D. W. Cheung. Mining frequent spatio-temporal sequential patterns. In *ICDM’05*, pages 82–89, 2005.
- [7] Y. Chen and L. Tu. Density-based clustering for real-time stream data. In *KDD’07*, pages 133–142, 2007.
- [8] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *KDD’11*, pages 1082–1090, 2011.
- [9] N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. In *Proceedings of the National Academy of Sciences*, volume 106, page 15274, 2009.
- [10] F. Giannotti, M. Nanni, D. Pedreschi, and F. Pinelli. Trajectory pattern mining. In *KDD’07*, pages 330–339, 2007.
- [11] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
- [12] Z. Huang and D. K. J. Lin. The time-series link prediction problem with applications in communication surveillance. *INFORMS Journal on Computing*, 21(2):286–303, 2009.
- [13] H. W. Lauw, E.-P. Lim, H. Pang, and T.-T. Tan. Stevent: Spatio-temporal event model for social network discovery. *ACM Transactions on Information Systems*, 28(3), 2010.
- [14] P.-R. Lei, T.-J. Shen, W.-C. Peng, and I.-J. Su. Exploring spatial-temporal trajectory model for location prediction. In *MDM’11*, pages 1082–1090, 2011.
- [15] X. Liu, Q. Hey, Y. Tiany, W.-C. Lee, J. McPherson, and J. Han. Event-based social networks: Linking the online and offline social worlds. In *KDD’12*, pages 1032–1040, 2012.
- [16] A. Monreale, F. Pinelli, and R. Trasarti. Wherenext: a location predictor on trajectory pattern mining. In *KDD’09*, pages 637–645, 2009.
- [17] M. Morzy. Mining frequent trajectories of moving objects for location prediction. In *MLDM’07*, pages 667–680, 2007.
- [18] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. Mining user mobility features for next place prediction in location-based services. In *ICDM’12*, pages 1038–1043, 2012.
- [19] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *WSDM’12*, pages 723–732, 2012.
- [20] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla. When will it happen?: relationship prediction in heterogeneous information networks. In *WSDM’12*, pages 663–672, 2012.
- [21] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabási. Human mobility, social ties, and link prediction. In *KDD’11*, pages 1100–1108, 2011.
- [22] G. Welch and G. Bishop. An introduction to the kalman filter. In *SIGGRAPH’01*, 2001.
- [23] Y. Zheng. Tutorial on location-based social networks. In *WWW 2012*, 2012.