Prediction of User Location Using the Radiation Model and Social Check-Ins*

Alexey Tarasov
Applied Intelligence Research Centre
School of Computing
Dublin Institute of Technology
Kevin St, Dublin 8, Ireland
aleksejs.tarasovs@student.dit.ie

Felix Kling
National Centre for Geocomputation
Iontas Building
NUI Maynooth
Maynooth, Co. Kildare, Ireland
felix.kling@nuim.ie

Alexei Pozdnoukhov
Department of Civil and Env. Engineering
University of California, Berkeley
Berkeley, CA 94720
alexei.pozdnoukhov@gmail.com

ABSTRACT

Location-based social networks serve as a source of data for a wide range of applications, from recommendation of places to visit to modelling of city traffic, and urban planning. One of the basic problems in all these areas is the formulation of a predictive model for the location of a certain user at a certain time. In this paper, we propose a new approach for predicting user location, which uses two components to make the prediction, based on (i) coordinates and times of user check-ins and (ii) social interaction between different users. We improve the performance of a state-of-the art model using the radiation model of spatial choice and a social component based on the frequency of matching check-ins of user's friends. Friendship is defined by the presence of reciprocal following on Twitter. Our empirical results highlight an improvement over the state-of-the-art in terms of accuracy, and suggest practical solutions for spatio-temporal and socially-inspired prediction of user location.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—data mining, spatial databases and GIS

Keywords

Location-based social networks, urban mobility, radiation model, social influence

*The source code that implements all the procedures described in this paper, as well as the dataset, is available at `https://github.com/alexeytarasov/spatial-temporal-social-model`.

UrbComp'13 August 11–14, 2013, Chicago, Illinois, USA.

1. INTRODUCTION

Location-based social networks—as well as locational services built into social networks—allow individuals to perform so-called *check-ins*, i.e. to voluntarily denote their current location and to share it with people they know. Such services are becoming more and more popular, as they provide a connection between virtual social life and real-world activities. For instance, Foursquare[1] allows users to share their experience with different venues such as restaurants or cinemas, and to get special awards called *badges* for logging many check-ins. Beyond traditional location-based social networks, social and spatial elements are becoming a popular feature in many services. An example is Bikely[2], a website for mapping and sharing cycling routes and mountain bike trails. Such networks can serve not only as entertainment, but also as a source of data for a wide range of applications. One of the basic problems in such applications is how to predict the location of a certain user at a certain time. For instance, solving this problem is essential to empower traditional location-based services operating on mobile devices [8], and gain knowledge for various areas including venue recommender systems [2], modelling of city traffic [9], and urban planning [6].

State-of-the-art approaches in prediction of user location in location-based social networks operate both on spatio-temporal (when and where the user checked in in the past) and social information (who are the user's social connections). This is the approach used by Periodic & Social Mobility Model (PSMM) [4]. However, the PSMM makes certain assumptions which are not absolutely valid for this task. For instance, discrete check-ins are modelled using continuous Gaussian distributions. However, selecting a travel destination (and performing a check-in at a selected venue) is essentially a discrete choice process [3]. In this paper, we focus on modelling a set of possible destinations with multinomial distributions in order to improve the performance of the PSMM. We propose a new approach which uses spatio-temporal information from user check-ins by applying the radiation model [14] instead of Gaussians used by PSMM.

[1]https://foursquare.com/
[2]http://www.bikely.com/

Also, we describe a novel approach to predict user location based on information about past check-ins from the user's social connections. We show that both approaches significantly outperform the PSMM baseline. In addition, our findings suggest that even if nothing is known about a user's own check-ins, social connections and their check-ins provide a significant amount of information. Previous work applied dynamic Bayesian networks to learn a user location from friends' locations [13]. Our social model alone was able to predict the user location with as high accuracy as the approach using both social and spatio-temporal information.

The paper is structured as follows. Section 2 describes related research including the PSMM and radiation model, while Section 3 describes the experiment methodology. The results are described and discussed in Section 4. Section 5 concludes the paper and suggests possible directions for future work.

2. RELATED WORK

One of the fundamental research questions in location-based social networks is prediction of user location. Many sources of location-based information are currently available, including cellphone usage logs [15], "check-ins" in location-based social network services [10] or geo-tagged Twitter texts [12, 13], as well as databases of high precision GPS traces [17]. There is some evidence that user location is highly correlated with his previous check-in activity and the activity of his social connections [16]. One of the most recent models for prediction of user location is the Periodic Mobility Model (PMM) [4]. While the model proposed in [16] does not take time into account, the PMM, in contrast, assumes that short-range travel, such as commuting from home to work or going to lunch, is periodic both spatially and temporally. The PMM uses the notion of two states which can be called *Home* and *Work*. While in *Work* state, an individual might check in at his workplace or around it, keeping the venues mostly work-related. At the same time, his mobility pattern in *Home* state is likely to be different: he should be in the vicinity of his home and might visit different entertainment-related venues. In other words, both *Home* and *Work* check-ins tend to be centred around a particular venue. The PMM consists of two parts: spatial and temporal. The spatial part of the PMM represents both clusters as 2D Gaussian distributions. The temporal part models the probability of the individual being in *Home* or *Work* state at a particular time of the day and is represented as a 1D Gaussian distribution. The PMM has eighteen parameters:

- Six temporal: mean and variance of time when the individual is in *Home* or *Work* state and two probabilities that any check-in belongs to either *Home* or *Work* state.
- Twelve spatial: two 2D means and two corresponding 2x2 covariance matrices.

These parameters are fitted using an expectation-maximisation (EM) algorithm. The probability that the individual checks in at venue x at time t is

$$\begin{aligned} P[x(t) = x] = P[x(t) = x|z(t) = H] \cdot P[z(t) = H] \\ + P[x(t) = x|z(t) = W] \cdot P[z(t) = W], \end{aligned}$$

where $P[z(t) = H]$ and $P[z(t) = W]$ are the temporal part (1D Gaussians). The expressions

$$\begin{aligned} &P[x(t) = x|z(t) = H] \text{ and} \\ &P[x(t) = x|z(t) = W] \end{aligned}$$

represent the spatial part (2D Gaussians fitted to *Home* and *Work* check-ins respectively).

Predictive modelling of human movement beyond regular periodic commutes, such as that modelled by PMM, is a challenging task. Empirical observations [1, 4, 13] demonstrate the social influence on the formation of atypical patterns of mobility. People tend to follow the recommendations of their friends in planning travel. They may also join them on a trip to explore new areas or visit particular places for recreation, leisure or tourism. Accommodating the social influence in such scenarios is also a recognised challenge within urban transportation design [5]. In this domain discrete choice models [3] are a generally accepted underlying methodology.

The creators of the PMM also argue that check-ins that cannot be properly modelled by it are mostly caused by social influence, i.e. visiting a friend or going with a friend to some distant location [4]. In order to model this behaviour, they offered the Periodic & Social Mobility Model (PSMM) which is an extension of PMM. PSMM consists of two components: the spatio-temporal component, which is in fact the PMM, and a social component. The first component is trained as described above. The second is based on the assumption that the check-ins that do not fit well with the PMM are the result of social activity. The social component of the PSMM is fitted to these check-ins. The main intuition behind the PSMM social component is that the probability of the individual checking in at a certain venue is high if many of his friends checked in at a neighbouring venue at approximately the same time. The probability of the individual performing a social check-in at time t to a venue with coordinates x is

$$P[x(t) = x|z(t) = S] \sim \sum_{(t_j,x_j)\in J} |t_j - t|^{-\alpha} \cdot \|x - x_j\|^{-\beta}, \quad (1)$$

where J is the set of check-ins made by the individual's friends on the same day. Each check-in $j \in J$ happened at time t_j at coordinates x_j. Parameters α and β are to be tuned using an EM algorithm. When the training of the PSMM model is complete, both the spatio-temporal and social components are used to calculate the probability of the individual checking in at the venue with coordinates x at time t as

$$\begin{aligned} P[x(t) = x] = P[x(t) = x|z(t) = H] \cdot P[z(t) = H] \\ + P[x(t) = x|z(t) = W] \cdot P[z(t) = W] \\ + P[x(t) = x|z(t) = S]. \end{aligned} \quad (2)$$

Unfortunately, the authors of the PSMM did not provide a great deal of detail on the fitting of parameters α and β. It is also unclear how the PMM distinguishes between social and non-social check-ins.

The task of predicting user location was also addressed by [11], who used the same information as PSMM, but also proposed to use knowledge about categories of venues (cinema, coffee shop, football stadium etc.) as well as global patterns of mobility, i.e. number of check-ins at venues performed by all users, not only social connections of a given user.

An alternative but untested approach to modelling check-ins is the radiation model [14] which was originally proposed as a way of modelling mobility and migration patterns. One of the main advantages of the radiation model is that it does not require tuning of parameters. The radiation model

is used to calculate the intensity of flow T_{ij} between locations i and j, located at distance r_{ij} from each other, having populations m and n respectively:

$$T_{ij} = T_i \cdot \frac{m \cdot n}{(m + s_{ij}) \cdot (m + n + s_{ij})},$$

where T_i is the total number of people leaving the location i and s_{ij} is the total population in the circle of radius r_{ij} centred at i (this number does not include the population of i and j). The illustrative example used in the original paper is job hunting: an individual moves from his home location i to some other location j to find a job as close to his home location as possible. It was assumed that the number of job opportunities strongly correlates with the population of the location. Thus, the individual will have a better chance of finding a job in a location with high n, and move there. Also, if the area around i is highly populated, it is probable that the individual will find some location, which is closer to i than j, but also has good employment possibilities. In order to account for that factor the radiation model uses s_{ij}.

Originally the radiation model was used to model large scale mobility patterns, i.e. movement from one U.S. state to another. However, [9] used the radiation model to simulate the movement of individuals between shops, restaurants and other venues in the Dublin area. It was shown that the probability of an individual moving from location i to location j is

$$P_{ij} = \frac{m \cdot n}{(m + s_{ij}) \cdot (m + n + s_{ij})},$$

where m and n represent capacities of venues rather then populations of a given location. This interpretation of the radiation model uses an assumption that individuals will be more attracted to venues with higher capacities; such venues are more likely to be attractive in the same way that highly populated areas are more likely to present more employment opportunities.

3. EXPERIMENT

This paper proposes a novel model of predicting user location. Its structure is similar to PSMM, as it also has two components. The first is based on spatio-temporal information and uses the radiation model, while the second operates on social information and uses an intuition similar to that used in PSMM. We test these components separately and compare them to the respective components of the PSMM. Thus, our experiments involve two spatio-temporal and two social components. We are interested in three research questions:

1. What is the best combination of spatio-temporal and social components?
2. What is the best social component?
3. What is the best spatio-temporal component?

This section is structured as follows. Section 3.1 describes the dataset we used, Section 3.2 explains in detail our implementation of PSMM. We then proceed with a description of our spatio-temporal (Section 3.3) and social (Section 3.4) components. We finish with a description of the experimental setup in Section 3.5.

3.1 Dataset

For our experiments, we used our own dataset of check-ins gathered in July 2012. The dataset consists of tweets, associated with check-ins that users made on Foursquare between 13/07/2012 and 24/07/2012. There are 90 users in the dataset, each having from 12 to 212 check-ins (31.13 check-ins on average). Each user has 7.48 social connections on average, with the minimum being 2 and maximum being 37. For each check-in, we captured latitude, longitude, date, time, venue ID, and tweet text associated with the check-in. To facilitate tweet text analysis we focused on users from English-speaking countries. The distribution of check-ins is shown in Figure 1.

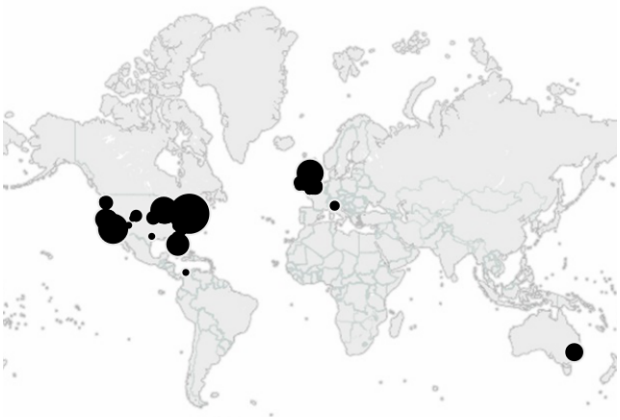

Figure 1: Distribution of check-ins from the dataset. Relatively big clusters of check-ins located in New York, Chicago, Miami, San Francisco, Los Angeles and Edinburgh.

This dataset was created in several steps, as follows. The check-in data was collected through the Twitter streaming API, filtering for tweets containing "4sq" keyword. Further tests were applied to ensure that each such tweet was indeed a check-in. Along with the message and check-in URL, each response contained information about the user on Twitter. To get a more specific location for the check-in, as well as information about the venue, we extracted data from the check-in URL and queried the Foursquare check-in API. Finally, to identify the social connections, we used Twitter's REST API and retrieved the friend lists for each user. We considered that two users were friends if both of them mutually followed each other.

3.2 Implementation of baselines

We implemented the PMM, the spatio-temporal component of the PSMM, as described in the original paper [4]. The process of PSMM fitting began by training the original PMM, as described in [4]. We considered that any check-in has a certain probability of being social. This probability was calculated according to Equation 1. These probabilities were normalised such that all values were bounded by $[0; 1]$. That allowed us to take $\alpha = \beta = 1$. Following values were calculated when the social component of the PSMM was fit-

ted to the collection of n check-ins $(t_i, x_i) \in I\,(i = 1, 2, ..., n)$:

$$\begin{aligned}
\Delta^t_{min} &= \min_{\substack{(t_i,x_i)\in I \\ (t_j,x_j)\in J_i}} |t_i - t_j| \\
\Delta^t_{max} &= \max_{\substack{(t_i,x_i)\in I \\ (t_j,x_j)\in J_i}} |t_i - t_j| \\
\Delta^x_{min} &= \min_{\substack{(t_i,x_i)\in I \\ (t_j,x_j)\in J_i}} \|x_i - x_j\| \\
\Delta^x_{max} &= \max_{\substack{(t_i,x_i)\in I \\ (t_j,x_j)\in J_i}} \|x_i - x_j\|,
\end{aligned} \tag{3}$$

where J_i is the set of check-ins by the user's social connections on the same day as t_i. Thus, our implementation of the PSMM had parameters Δ^t_{min}, Δ^t_{max}, Δ^x_{min}, Δ^x_{max} instead of α and β. After this trivial fitting was complete, the PSMM was ready to calculate the probability of the check-in (t, x) being social in the following way:

$$P[x(t) = x | z(t) = S] \propto \left(\sum_{(t_j,x_j)\in J} \frac{|t - t_j|}{|\Delta^t_{max} - \Delta^t_{min}|} \cdot \frac{\|x - x_j\|}{\|\Delta^x_{max} - \Delta^x_{min}\|} \right)^{-1},$$

where J is the set of check-ins by user's social connections on the same day as t. Both multipliers under the sum can take a value in $[0; 1]$, thus, each addend of the sum will also be in $[0; 1]$. This means that the sum can have values bigger than 1. To interpret it as probability, we normalised this sum by maximal and minimal values of $|t_j - t| \cdot \|x - x_j\|$ in a manner similar to Equation 3.

Although this methodology draws heavily from the procedure described in [4], the authors omitted some important implementation details. As a result, our implementation might differ slightly from the original.

3.3 Spatio-temporal component

We handle time in the same manner as PMM, i.e. we use 1D Gaussians. However, we calculate spatial probability in a different way: instead of Gaussians, we use a radiation model. Separate radiation models are trained for *Home* and *Work* check-in clusters. We assume that in each cluster there is one *central venue*, a venue at which the user checks in very often and from which the user usually moves to different venues within the same cluster. A building where the user works is a good example of a central venue for the *Work* cluster. For a given cluster, we consider the probability of checking in at venue X to be equal to the probability of the user moving from the central venue to venue X. For the sake of simplicity, we assume that movements occur only from the central venue to other venues. Our interpretation of parameters m and n is similar to [9]. The n parameter is the attractiveness of the venue, and we consider it to be equal to the total number of check-ins made at that venue by all 90 users in the dataset. The m parameter is on the same scale as n, but is used only for *Home* and *Work* central venues. It denotes the attractiveness of the central venue. A user would rather stay at the central venue if it has a high m value. Thus, if the central venue is more attractive than other venues in the same cluster, it is unlikely that a user will move from the central venue.

The spatio-temporal component training algorithm (Algorithm 1) starts by assigning all check-ins to *Home* and *Work* clusters in a random fashion (line 1) and estimating m for each cluster as a mean popularity score of venues from that cluster (line 2). At the E-step of the algorithm (lines 4-8) the parameters of the model are re-evaluated as

Algorithm 1: Algorithm for training the spatio-temporal component based on the radiation model (see Section 3.3)

Data:
check_ins (set of all user's check-ins)
n (popularity scores of all venues, where the user has checked in)
Result:
σ_H, σ_W: circular standard deviation of time when the user checks in at home and work venues
τ_H, τ_W: circular mean of time when the user checks in at home and work venues
P_{c_H}, P_{c_W}: proportions of home and work check-ins
$central_H, central_W$: central home and work venues
m_H, m_W: m values for home and work
// `RandomAssignment (X)`: divides venues X into two clusters randomly
// `MaxLikelihood (H, W)`: calculates $\sigma_H, \sigma_W, \tau_H, \tau_W$ parameters via maximum likelihood estimation for home check-ins H and work check-ins W
// `Time (X)`: extracts time from the check-in X
// `Coordinates (X)`: extracts coordinates (latitude and longitude) from the check-in X
// `Venue (X)`: extracts venue(s) from the check-in(s) X
// `CentralVenue (X, n, m)`: returns a central venue from the set of venues X, having popularity scores n, given m, the popularity of the current central venue
// `OptimiseM (m, centre, n, venues)`: returns a new value of m performing a gradient descent optimisation, considering that there are venues *venues* with popularity scores n and the venue *centre* is the central one
// `PRadiation (venue, central_venue, m, venues, n)`: returns the probability that the user will check in at the venue *venue* given a central venue, m value, list of all venues and popularity scores of all venues

1 $H, W \leftarrow$ `RandomAssignment`$(check\_ins)$;
2 $m_H \leftarrow \frac{\sum_{check\_in\in H} n(check\_in)}{|H|}$; $m_W \leftarrow \frac{\sum_{check\_in\in W} n(check\_in)}{|W|}$;
3 **repeat**
4 $\quad \sigma_H, \sigma_W, \tau_H, \tau_W \leftarrow$ `MaxLikelihood`(H, W);
5 $\quad P_{c_H} \leftarrow \frac{|H|}{|H|+|W|}$; $P_{c_W} \leftarrow \frac{|W|}{|H|+|W|}$;
6 $\quad central_H \leftarrow$ `CentralVenue(Venue`$(H), n, m_H)$;
7 $\quad central_W \leftarrow$ `CentralVenue(Venue`$(W), n, m_W)$;
8 $\quad m_H \leftarrow$ `OptimiseM`$(m_H, central_H, n,$ `Venue`$(H))$;
$\quad\quad m_W \leftarrow$ `OptimiseM`$(m_W, central_W, n,$ `Venue`$(W))$;
9 $\quad$ **foreach** *venue in* `Venue` $(H \bigcup W)$ **do**
10 $\quad\quad p_{spatial_H}(venue) \leftarrow$ `PRadiation`$(venue, central_H,$ `Venue`$(H \bigcup W), n)$;
11 $\quad\quad p_{spatial_W}(venue) \leftarrow$ `PRadiation`$(venue, central_W,$ `Venue`$(H \bigcup W), n)$;
12 $\quad \Sigma_H \leftarrow \sum_{venue\in \text{Venue}(H \bigcup W)} p_{spatial_H}(venue)$;
13 $\quad \Sigma_W \leftarrow \sum_{venue\in \text{Venue}(H \bigcup W)} p_{spatial_W}(venue)$;
14 $\quad$ **foreach** *venue in* `Venue` $(H \bigcup W)$ **do**
15 $\quad\quad p_{spatial_H}(venue) \leftarrow p_{spatial_H}(venue)/\Sigma_H$;
16 $\quad\quad p_{spatial_W}(venue) \leftarrow p_{spatial_W}(venue)/\Sigma_W$;
17 $\quad\quad p_{temporal_H} \leftarrow 0$; $p_{temporal_W} \leftarrow 0$;
18 $\quad\quad$ **foreach** *check_in in check-ins in venue* **do**
19 $\quad\quad\quad t \leftarrow$ `Time`$(check\_in)$;
20 $\quad\quad\quad N_H \leftarrow \frac{P_{c_H}}{\sqrt{2\pi\sigma_H^2}} \exp(-(\frac{\pi}{12})^2 \frac{(t-\tau_H)^2}{2\sigma_H^2})$;
21 $\quad\quad\quad N_W \leftarrow \frac{P_{c_W}}{\sqrt{2\pi\sigma_W^2}} \exp(-(\frac{\pi}{12})^2 \frac{(t-\tau_W)^2}{2\sigma_W^2})$;
22 $\quad\quad\quad p_{temporal_H} \leftarrow p_{temporal_H} + \frac{N_H}{N_H+N_W}$;
23 $\quad\quad\quad p_{temporal_W} \leftarrow p_{temporal_W} + \frac{N_W}{N_H+N_W}$;
24 $\quad\quad p_{temporal_H}(venue) \leftarrow p_{temporal_H}/|check\_ins|$;
25 $\quad\quad p_{temporal_W}(venue) \leftarrow p_{temporal_W}/|check\_ins|$;
26 $\quad\quad p_H(venue) \leftarrow p_{spatial_H} \cdot p_{temporal_H}$;
27 $\quad\quad p_W(venue) \leftarrow p_{spatial_W} \cdot p_{temporal_W}$;
28 $\quad H_{old} \leftarrow H$; $W_{old} \leftarrow W$; $H \leftarrow \emptyset$; $W \leftarrow \emptyset$;
29 $\quad$ **foreach** *check_in in* $H_{old} \bigcup W_{old}$ **do**
30 $\quad\quad$ **if** $P_H(check\_in) > P_W(check\_in)$ **then**
31 $\quad\quad\quad H \leftarrow H \bigcup check\_in$;
32 $\quad\quad$ **else**
33 $\quad\quad\quad W \leftarrow W \bigcup check\_in$;
34 **until** $H = H_{old}$;
35 **return** $\sigma_H, \sigma_W, \tau_H, \tau_W, P_{c_H}, P_{c_W}$;

the contents of the *Home* and *Work* clusters have changed since the previous iteration. At the M-step (lines 9-27) the calculation of new probability values for all venues takes place: spatial probabilities are calculated (lines 10-11) and normalised (lines 12, 13, 15 and 16). The same happens with the temporal probabilities (lines 17-23). The temporal probability for a certain venue is a mean temporal probability of all check-ins made at this venue (lines 24-25). The spatio-temporal probability for each venue is calculated as a product of corresponding spatial and temporal probabilities (lines 26-27). After spatio-temporal probabilities are calculated, the venues are re-assigned to *Home* and *Work* clusters (lines 28-33). The algorithm finishes if the assignment to *Home* and *Work* clusters has not changed since the previous iteration (line 34). Algorithm 2 explains how the central venue is determined, while Algorithm 3 describes how the spatial probability of checking in at particular venue is performed.

When training is finished, the probability that the user will check in at a certain venue happens as follows. First, the temporal probability has to be calculated (similar to lines 14-16 of Algorithm 1) to determine which—*Home* or *Work*—radiation model to use for prediction. Then Algorithm 3 is used to calculate probabilities for all candidate venues and the venue with the biggest probability is returned as the predicted result.

Algorithm 2: CentralVenue: algorithm for finding a central venue

Data:
venues: list of all venues
n: capacities of all venues from *venues*
m: current value of m
Result: central venue
// `GetVenuesInsideCircle (centre, venue, venues):` calculates an area bounded by a circle whose radius is the distance between *centre* and *venue* and returns all elements from *venues* that are inside this circle. The set does not include *centre* and *venue* venues
// `Set (venues):` removes duplicates from *venues*

```
1  venues ← Set(venues);
2  foreach venue in venues do
3      foreach j in venues do
4          venues_nearby ← GetVenuesInsideCircle(venue, j);
5          if |venues_nearby| > 0 then
6              S(venue) ← Σ_{i∈venues_nearby} n(i);
7          else
8              if j = venue then
9                  S(venue) ← 0;
10             else
11                 continue;
12         p_j = m·n(j) / ((m+n(j))(m+n(j)+S(j)));
13     likelihood ← 0;
14     foreach i in venues do
15         likelihood ← likelihood + log (m·n(i)) / ((m+S(i))·(m+S(i)+n(i)))
16     likelihoods(venue) ← likelihood;
17 return arg max_{venue∈venues} likelihoods(venue)
```

3.4 Social component

Our social component makes use of check-ins by socially connected users, performed only at the same venue, not at all venues as in PSMM. Let V be the set of venues where a user has checked in at least once. In order to get the social probability $P[x(t) = x|z(t) = S]$ of the user checking in at venue $x \in V$ at time t, we conducted the following

Algorithm 3: PRadiation: algorithm for the calculation of the probability of check-in at *venue* using the radiation model

Data:
venue: target venue
central_venue: central venue
venues: list of all venues
n: capacities of all venues from *venues*
m: current value of m
Result: probability of check-in at the target venue *venue*
// `GetVenuesInsideCircle (centre, venue, venues):` calculates an area bounded by a circle whose radius is the distance between *centre* and *venue* and returns all elements from *venues* that are inside this circle. The set does not include *centre* and *venue* venues

```
1  nearby ← GetVenuesInsideCircle(central_venue, venue, venues);
2  if |nearby| > 0 then
3      S ← Σ_{i∈nearby} n(i);
4  else
5      S ← 0;
6  return (m·n(venue)) / ((m+S)(m+S+n(venue)));
```

procedure:

1. For each venue $v \in V$ the number of social check-ins N_v was calculated. Social check-ins were check-ins (i) which were made by people who have a social connection with the user and (ii) which took place plus/minus two hours from time t.

2. Each N_v was divided by $\max_{v \in V} N_v$, after which $N_v \in [0; 1], v \in V$.

3. N_x was the social probability.

3.5 Experimental setup

The experiment was conducted separately for each of the 90 users in our dataset, using check-ins made from Mondays through Thursdays in order to focus on the check-in behaviour exhibited on a typical weekday. For each user, seven models corresponding to the seven approaches given in Table 1 were trained. Each approach represents a different combination of spatio-temporal and social components, using Equation 2 to calculate probabilities. If a spatio-temporal or social component was absent, the corresponding probability was zero.

Prediction accuracy was used as a performance measure. It represents the proportion of cases where the model was able to correctly predict a check-in venue given a time of day. Five-fold cross validation was used in our experiments.

As all models contain a random component (i.e. the initial assignment of check-ins to *Home* and *Work* clusters), we trained 10 instances of each model for each user and used the model with the highest likelihood for prediction. In order to take different distributions of check-ins into folds into account, we conducted all experiments 10 times, and reported average accuracies.

To measure the significance of any performance differences between approaches, we used Bergmann-Hommel's procedure [7]. This procedure divides all approaches into a few ranked groups. If two approaches are in the same group, there is no statistically significant difference between their performance. In contrast, if two approaches belong to different groups, the difference between them is significant. When

Table 1: Description of approaches used to predict user location

Approach	spatio-temporal model	Social model
Stanford+None	Periodic Mobility Model [4]	None
Stanford+Stanford	Periodic Mobility Model [4]	Periodic & Social Mobility Model [4]
Stanford+Matching	Periodic Mobility Model [4]	Matching check-ins of user's friends (Section 3.4)
Radiation+None	Radiation model (Section 3.3)	None
Radiation+Stanford	Radiation model (Section 3.3)	Periodic & Social Mobility Model [4]
Radiation+Matching	Radiation model (Section 3.3)	Matching check-ins of user's friends (Section 3.4)
None+Matching	None	Matching check-ins of user's friends (Section 3.4)

there was a need to compare the performance of two approaches, we used the Wilcoxon test.

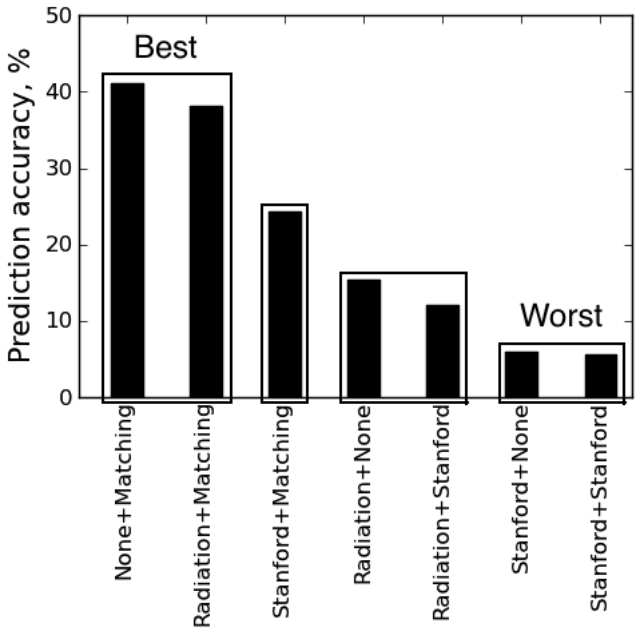

Figure 2: Comparison of accuracies of different approaches to predicting user location. Rectangles denote groups detected by Bergmann-Hommel's procedure (the difference in accuracy between approaches in the same group is statistically insignificant).

4. RESULTS AND DISCUSSION

Approaches which used Gaussian mixtures as a spatio-temporal component (*Stanford+None*, *Stanford+Stanford* and *Stanford+Matching*) were unable to converge on 29 out of 90 users (32% of the total user count). One possible explanation is that at some iteration of the EM algorithm all check-ins for either *Home* or *Work* cluster were very close to each other, or even were made to the same venue. It resulted in numerical difficulties while calculating the covariation matrix of check-in latitudes and longitudes. When this arose, we assumed that the accuracy of the corresponding model was zero. The average accuracies of all seven approaches we used are depicted in Figure 2, where rectangles denote groups detected by Bergmann-Hommel's procedure.

Here are the results for each research question from Section 3:

1. **What is the best combination of spatio-temporal and social components?** *None+Matching*, the approach which operates solely on social connection information, proved to be the most accurate. It suggests that social connections alone carry a lot of information about user location, even when the prediction model is very simple. In our experiments we often encountered a situation where, for a particular user at a certain time, two or more venues had the same number of check-ins made by friends. This means that the social probabilities of checking in at such venues for that user at that time were equal. A spatio-temporal component helped to resolve these ties. When the radiation model was used as a spatio-temporal component (*Radiation+Matching*), no statistically significant benefit was discovered. The use of Gaussian mixtures (*Stanford+Matching*) disimproved prediction accuracy.

2. **What is the best social component?** As illustrated in Figure 2, *Radiation+Stanford* and *Radiation+None* were in the same group. *Stanford+Stanford* and *Stanford+None* were also grouped together. This indicates no statistically significant increase in accuracy when the PSMM social component was used, compared to using just a spatio-temporal component. This conclusion is supported by the original paper [4] which identified only a small difference, although no statistical testing was carried out.

3. **What is the best spatio-temporal component?** If no social information is available, only two methods can work: *Stanford+None* and *Radiation+None*. *Radiation+None* was able to correctly predict a venue in 15.46% of cases, while *Stanford+None* was approximately one third as accurate (5.88%). The *Radiation+None* approach proved to be significantly better (Wilcoxon test p-value $= 2.49 \cdot 10^{-10}$). This suggests that the radiation model is indeed capable of predicting user locations and is more suited to this task than the Gaussian mixture model used in the PSMM spatio-temporal component.

5. CONCLUSIONS AND FUTURE WORK

The results of our experiment strongly suggest that both the spatio-temporal and social components proposed in this paper perform better than the state-of-the-art baseline. The radiation model proved to be better than PMM, the baseline approach using Gaussian mixture models. The radiation model was not susceptible to the numerical difficulties which

prevented the training of the PMM in 32% of cases. Also, a very simple social component based on counting a number of check-ins by friends showed higher accuracy than a much more sophisticated PSMM baseline. This simple social component was also the most accurate, even by itself, without using any spatio-temporal information.

One possible direction for future work is to reproduce our results using more datasets. Also, conducting experiments involving more baseline methods might be worthwhile. For instance, it could be very promising to incorporate information about the venue type, similar to the approach by [11], in our model. In addition, it might be interesting to look into ways in which both the spatio-temporal and social components proposed in this paper can be further improved.

6. ACKNOWLEDGMENTS

The research presented in this paper was funded by Marie-Curie ITN GEOCROWD, a Strategic Research Cluster grant No. 07/SRC/I1168 and the 11/RFP.1/CMS/3247 award by Science Foundation Ireland under the National Development Plan. Alexey Tarasov was funded by Science Foundation Ireland grant No. 09/RFP/CMS/253.

The authors would like to thank Mr Stephen Kidney for help proofreading the paper.

7. REFERENCES

[1] K. W. Axhausen. Social networks, mobility biographies, and travel: survey challenges. *Environment and Planning B: Planning and Design*, 35:981–996, 2008.

[2] J. Bao, Y. Zheng, and M. Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pages 199–208, 2012.

[3] M. Ben-Akiva and S. R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. The MIT Press, 1985.

[4] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2011.

[5] E. R. Dugundji, A. Páez, T. A. Arentze, J. L. Walker, J. A. Carrasco, F. Marchal, and H. Nakanishi. Transportation and social interactions. *Transportation Research Part A: Policy and Practice*, 45(4):239 – 247, 2011. Special Issue: Transportation and Social Interactions.

[6] V. Frias-Martinez, V. Soto, H. Hohwald, and E. Frias-Martinez. Characterizing urban landscapes using geolocated tweets. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 239–248, 2012.

[7] S. Garcia and F. Herrera. An extension on "Statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, 9:2677–2694, 2008.

[8] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. Do, O. Dousse, J. Eberle, and M. Miettinen. The mobile data challenge: Big data for mobile computing research. In *Title*, Newcastle, UK, 2012.

[9] G. McArdle, A. Lawlor, E. Furey, and A. Pozdnoukhov. City-scale traffic simulation from digital footprints. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing (UrbComp)*, 2012.

[10] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: universal patterns in human urban mobility. *arXiv:1108.5355v4 [physics.soc-ph]*, 2011.

[11] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. Mining user mobility features for next place prediction in location-based services. In *Proceedings of IEEE International Conference on Data Mining, ICDM '12*, pages 1038–1043, 2012.

[12] A. Pozdnoukhov and C. Kaiser. Space-time dynamics of topics in streaming text. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '11, pages 8:1–8:8, New York, NY, USA, 2011. ACM.

[13] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 723–732, New York, NY, USA, 2012.

[14] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, Feb. 2012.

[15] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, Feb. 2010.

[16] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 325–334, New York, NY, USA, 2011. ACM.

[17] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W. Y. Ma. Understanding mobility based on GPS data. In *Proceedings of the 10th international conference on Ubiquitous computing*, UbiComp '08, pages 312–321, New York, NY, USA, 2008. ACM.