

Friendship and Mobility: User Movement In Location-Based Social Networks

Eunjoon Cho *
Stanford University
ejcho@stanford.edu

Seth A. Myers *
Stanford University
samyers@stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

ABSTRACT

Even though human movement and mobility patterns have a high degree of freedom and variation, they also exhibit structural patterns due to geographic and social constraints. Using cell phone location data, as well as data from two online location-based social networks, we aim to understand what basic laws govern human motion and dynamics. We find that humans experience a combination of periodic movement that is geographically limited and seemingly random jumps correlated with their social networks. Short-ranged travel is periodic both spatially and temporally and not effected by the social network structure, while long-distance travel is more influenced by social network ties. We show that social relationships can explain about 10% to 30% of all human movement, while periodic behavior explains 50% to 70%. Based on our findings, we develop a model of human mobility that combines periodic short range movements with travel due to the social network structure. We show that our model reliably predicts the locations and dynamics of future human movement and gives an order of magnitude better performance than present models of human mobility.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications – *Data mining*

General Terms: Algorithms, theory, experimentation.

Keywords: Human mobility, Communication networks, Social networks.

1. INTRODUCTION

While we would like to believe that our movement and mobility patterns have a high degree of freedom and variation, at a global scale human mobility exhibits structural patterns subject to geographic and social constraints. One would expect that people exhibit strong periodic behavior in their movement as they move back and forth between their homes and workplaces [7, 23]. Mobility is probably also constrained geographically by the distance one can travel within a day [12]. Moreover, mobility may further be shaped by our social relationships [11, 26] as we may be more likely to visit places that our friends and people similar to us visited in the past.

* Authors contributed equally to this work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD'11, August 21–24, 2011, San Diego, California, USA.

Copyright 2011 ACM 978-1-4503-0813-7/11/08 ...\$10.00.

Even though the above are some of the most fundamental questions and hypotheses about the dynamics of human mobility, answers to them remain largely unknown mostly due to the fact that reliable large scale human mobility data has been hard to obtain. Recently, however, location-based online social networking applications have emerged, where users share their current location by *checking-in* on websites such as Foursquare, Facebook, Gowalla, etc. While traditionally records of calls made by cell phones have been used to track the location of the cell phone towers associated with the calls [12, 17, 30], location-based social networks [15, 29, 28] provide an important new dimension in understanding human mobility. In particular, while cell phone data provides coarse location accuracy, location-based social networks provide location-specific data, as one can distinguish between a check-in to the office on the 2nd floor and a check-in to a coffee shop on the 1st floor of the same building. On the other hand, check-ins to location-based social networks are usually sporadic [28] while cell phone data provides better temporal resolution as a user “checks-in” whenever she makes or receives a call. Both types of data also contain network information. Location-based social networks maintain explicit friendship networks, while in mobile phones the network can be inferred from the communication network [8].

This data allows for studying the three main aspects of human mobility: geographic movement (where do we move?), temporal dynamics (how often do we move?) and the social network (how do social ties interact with movement?). Previous studies have mostly focused on at most two out of these three aspects. For example, research has focused on building models of human movement and dynamics [2, 12, 30], understanding the role of geographic distance and social interaction [1, 11, 24, 26, 31] and the change of our perception of (media) space [15, 4]. Studies have also explored how social networks are embed into the underlying geography [22, 19, 29], and how geo-data can be used to infer social ties [5, 8, 6]. In contrast, we study the interaction of all three aspects of human mobility to devise a coherent model of human movement and dynamics.

In the broader context, understanding and modeling human mobility has many applications and consequences. Knowledge of users’ locations can help improve large scale systems, such as cloud computing, content-based delivery networks [21], and location-based recommendations [13, 32, 33]. More broadly, accurate models of human mobility are essential for urban planning [10], understanding human migration patterns [14], and spread of diseases [9].

Present work. We study the relation between human geographic movement, its temporal dynamics, and the ties of the social network. In particular, we analyze the role of geography and daily routine on human mobility patterns as well as the effect of social ties, i.e., friends that one travels to meet. We seek to identify the

fundamental factors that govern human mobility and aim to answer questions, such as, how likely is a person going to a place because she has a friend there? How likely is a person to make a new friend by going to a place? How does this probability increase or decrease when she travels long distances from home?

Overview of results: Empirical findings. We use the data from two popular online location-based social networks, Gowalla and Brightkite, as well as a trace of 2 million mobile phone users from a European country. We observe that people generally move periodically within a bounded region but occasionally travel long distance. People are generally more likely to visit a distant place if it is in proximity of an existing friend. Thus, perhaps surprisingly, the short-ranged travel is less impacted by the social network structure, while if a person travels a long distance then they are more likely to travel near an existing friend. Overall, we find that the influence of friendship on one’s mobility is two times stronger than the influence of mobility on creating new friendships. We also note strong agreement and robustness of the patterns of human mobility between the cell phone and the location-based social network check-in data.

Generally, there are both benefits and limits to using mobility of friends to predict an individual’s location. We find that users are most likely to check-in right after a friend has checked-in to the same place, and the probability drops off following a power law as the time difference increases. We show that a similarity of movement trajectories is a strong indication of a tie in the social network. Despite strong correlation between friendship and mobility, there are limits in using friendship alone to predict mobility. For example, 84% of the people have less than 20% of their check-ins visited by a friend prior to their own check-in. Overall, social relationships can explain about 10% of human movement in cell phone data and up to 30% of movement in location-based social networks, while periodic movement behavior explains about 50% to 70%.

Overview of results: The mobility model. Building on our empirical findings we develop a *Periodic & Social Mobility Model* for predicting mobility of individuals. In particular, we build on the observation that people show strong periodic behavior throughout certain periods of the day alternating between primary (e.g., “home”) and secondary (e.g., “work”) locations on weekdays, and “home” and social network driven locations on weekends. Our model has three components: (1) a model of spatial locations that a user regularly visits, (2) a model of temporal movement between these locations, and (3) a model of movement that is influenced by the ties of the social network (i.e., meeting friends). We capture the temporal dynamics of transitioning between these locations with a day-specific periodic transition model. We model user locations with a mixture of Gaussians centered at “home” and “work” locations. On top of this we use a model of “social” movement that governs user behavior over the weekends and weeknights.

We develop an expectation-maximization based parameter estimation method and evaluate the predictive power of the *Periodic & Social Mobility Model* using three evaluation metrics. Experiments show that our model outperforms current mobility models for more than a factor of two on all three metrics. In particular, our model predicts the exact user location at any time with 40% accuracy, with an average relative distance error of 0.23% for cell phone locations, and 2.7% for social check-ins. We observe strong robustness of the model and agreement of results between the cell phone and the location-based social network check-in data.

Further related work. Models of human mobility consider movement either as a diffusive process [2], or a stochastic process centered about a single fixed point [12]. Our model considers human mobility as a time-varying stochastic process around several

fixed points. This additional flexibility of our model leads to a factor of two better predictive accuracy. More specialized mobility models have also been considered for wireless networks to model user transition between wireless access points [18, 27]. Similarly, there have also been attempts to capture the periodicity of human mobility using GPS position traces [20, 33] which are mostly governed by the physical embedding of road networks [16]. While GPS and wireless networking data allow for constantly tracing the user location, such studies have been limited to a relatively small number of users and small geographic areas (e.g., university campus). In contrast, our cell phone data covers two million users of a large country and check-ins from location-based social networks span the entire planet.

2. CHARACTERISTICS OF CHECK-INS

We use three different datasets that capture human mobility. First we consider two online location-based social networks. We collected all the public check-in data between Feb. 2009 and Oct. 2010 for Gowalla and Apr. 2008 to Oct. 2010 for Brightkite. The total number of check-ins for Gowalla is 6.4 million and 4.5 million for Brightkite. Gowalla and Brightkite also contain an explicit social network. In Gowalla the friendships are undirected and in Brightkite they are directed. For simplicity, we consider Brightkite as an undirected friendship network by only considering bi-directional edges. There are 196,591 nodes, 950,327 edges in Gowalla and 58,228 nodes, 214,078 edges in Brightkite.

To ensure that our observations on human movement are not specific to data based on check-ins from location-based social networks, we also include a dataset of cell phone location trace data. The cell phone dataset comes from a major cell phone service provider in Europe, and consists of nearly two million users and 450 million phone calls users over the course of 455 days. For each call, the nearest cell phone tower of both the person making and receiving the call was recorded. Overall, this means we have nearly 900 million “check-ins” with a spatial accuracy of about 3km. We only consider check-ins that happened within the country, and we create social network ties between pairs of people that have both called each other at least five times (10 calls total), yielding a network on 2 million nodes and 4.5 million edges.

In the remainder of the paper we will use word “check-in” to refer to an event when the time and the location of a particular user is recorded. For location-based social networks, this means that a user checked-in to a specific location using the online social network website/application, and for cell phone data this means a user either initiated or received a phone call. In all our plots we show Brightkite data in blue, Gowalla in red, and cell phone data in green.

Check-in behavior of users. In the following sections we analyze spatial and social characteristics of user check-ins. We are interested in understanding how far from their homes people tend to travel and how likely are they to meet social network friends at locations that they travel to. As user home locations are not explicitly given, we infer them by discretizing the world into 25 by 25km cells and defining the home location as the average position of check-ins in the cell with the most check-ins [29]. Manual inspection shows that this infers home locations with 85% accuracy.

First, we measure how far from their homes users tend to travel. Figure 1 plots the number of check-ins as a function of the distance from home for Brightkite, Gowalla and the cell phones. We observe that the distributions follow a power law with exponential cutoff in which there is an interesting kink at around 100km. The distribution decays faster for the travel of less than 100km and then flattens out for long distance jumps. We fit the power law parameters using

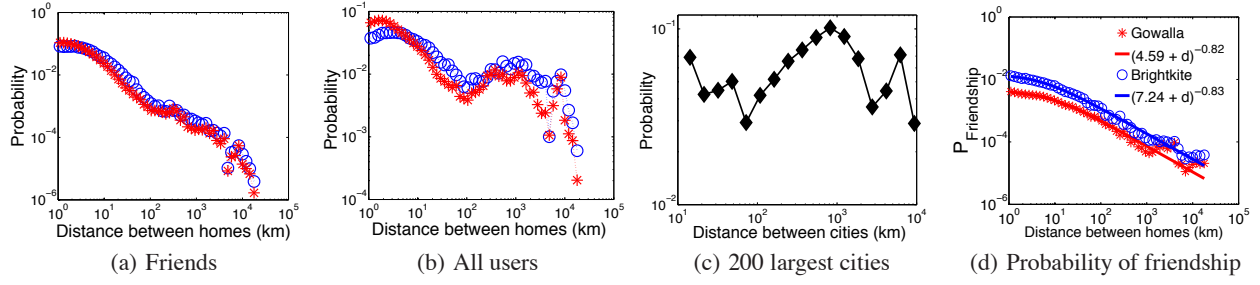


Figure 2: Distribution of home distances between (a) friends and (b) all users. (c) Distance between 200 world’s largest cities. (d) Probability of friendship as a function of distance when one accounts for non-uniform population density.

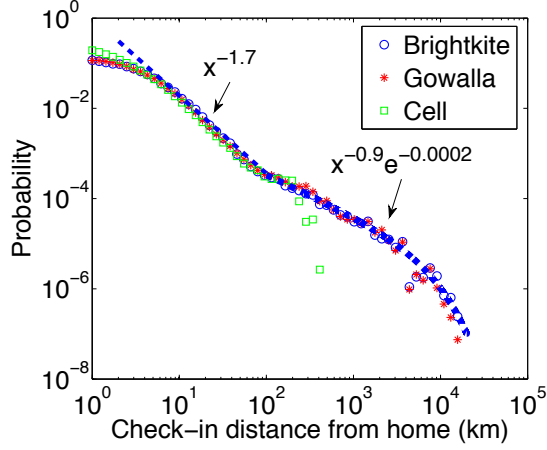


Figure 1: Fraction of check-ins as a function of distance traveled from home. Note the change in slope at around 100km.

maximum likelihood [3] and note a clear distinction in the power law exponents (1.7 vs 0.9) at around the 100km mark. Another interesting observation is that the distributions are extremely similar for all datasets. While Brightkite and Gowalla include check-ins from the whole world, cell phone data drops off quicker due to the small size of the country.

The distinct change of the distribution at the 100km mark can be explained by the non-uniform population density. We calculate the distribution of distances between the homes of friends in Fig. 2(a) and observe a similar kink at around 100km, which shows that the probability of two friends living a certain distance away decreases quickly at first but then slows down after the distance between the homes increases above 100km. Moreover, if we examine the distribution of distances between homes of random pairs of users in Fig. 2(b) we note an even stronger change around the 100km mark. As the number of pairs of people living closer than 100km decreases quickly with the distance, the distribution increases/flattens after the 100km mark. This abrupt change at 100km can be explained by the fact that users are geographically non-uniformly spread over the Earth and that humans cluster in cities. Interestingly, this suggests that around 100km is the typical human radius of “reach” [26] as it takes about 1 to 2 hours to drive such distance. For example, Fig. 2(c) shows distribution of distances between 200 of the world’s largest cities. The distribution of city distances follows a qualitatively similar pattern as the distances between pairs of users. We correct for this geographic non-uniformity and calculate the probability that a user has a friend at particular distance in Fig. 2(d). Now the kink at 100km disappears and the probability of friendship drops as a power law with an exponent of around 0.85 [1, 24]. This is interesting as it demonstrates that the kink at 100km is not due to some inherent property of our social interactions but is rather the

effect of non-uniform population density, where human population clusters in circles that are more than 100km apart.

3. FRIENDSHIP AND MOBILITY

So far, we examined how far from their homes people tend to travel and investigated the presence of the kink at the 100km mark. Now, we focus on the interaction of the person’s social network structure and their mobility.

Moving close to a friend’s home. We begin by investigating the sociability of human movement by measuring how likely is person A to travel close to the home of her friend B . We aim to understand how the location of A ’s friend B affects movement of A . We examine the fraction of check-ins that are in the vicinity of B ’s home. Intuitively we expect that people are more likely to move to a place in which they have friends, and that this likelihood decreases as the distance of travel increases. So far we saw that most of our friends live geographically close to us, and thus we would expect that they impact our movement the most. However, as we will see later, this is not the case.

To quantify this effect we proceed as follows. We say that user A “visits” her friend B if A checks-in within radius r of B ’s home, and we aim to compute $P_{data}(d)$, which measures the probability that A visits a friend given that A travels distance d from home. Quantity $P_{data}(d)$ is simply the fraction of A ’s check-ins at distance d from her home that occur in radius r (we set $r = 25km$) of one of A ’s friends. We experimented with various values of r and observed consistent behaviors.

Solid lines in Figure 3(a) plot $P_{data}(d)$, the fraction of friend visiting check-ins as a function of the distance traveled from home. Notice that both Gowalla and Brightkite exhibit similar behavior in that the probability of visiting a friend’s home levels off to a value of 0.3 after the 100km mark. This means that if a user travels more than 100km from her home, then there is a 30% chance that they will jump close to an existing friend’s home. Moreover, we observe that the probability of visiting a friend’s home remains constant after the 100km mark. The number of possible locations one can visit increases with the distance, and the number of friends decreases with the distance as well. This suggests that the probability of visiting a friend would decrease with the distance traveled (more possible locations to visit and less friends, and thus smaller probability of visiting a friend). Interestingly, we observe that the probability of visiting a friend remains constant as a function of distance traveled. We consider this surprising as it suggests that the effect friends have on our movement grows with their distance from us. To more precisely establish this we next compare the effect of friends on our movement to a null model that ignores the social network structure, and we then establish a causal relationship between the social network and mobility.

We compare the fraction of friend visiting check-ins $P_{data}(d)$ to a null model, $P_{null}(d)$, which quantifies the probability that a

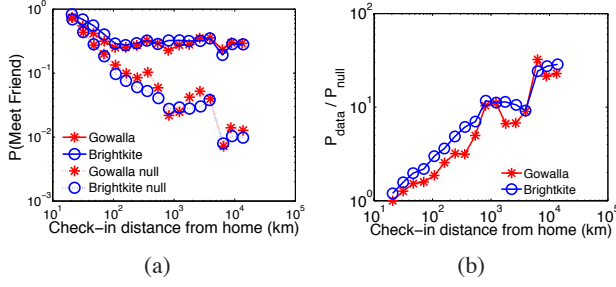


Figure 3: (a) Probability that a user will travel to friend's home as a function of distance traveled. (b) Influence of a friend relative to the null model.

friend will be within the vicinity of a user if the user would travel “randomly”. We consider the case where the user moves randomly with respect to the overall population density and measure the probability that a user travels to the vicinity of her friend's home. We plot $P_{\text{null}}(d)$ in Figure 3(a) using dashed lines, and (as expected) we observe a strong monotonic decrease. This means if people traveled independent of the network structure (but according to the overall population density) then the farther away they move from home the less likely they are to visit a friend's home. Additionally, in Fig. 3(b) we compare the real data with the null model to quantify the relative effect of friends on a user's movement. We observe that the actual influence of a friend on a user making a long distance jump increases with the distance. For example, the relative influence of a friend who lives 1,000km away is 10 times greater than the influence of a friend who lives 40km away. We also considered a version of the null model where users travel the exact same distance from home as in the real data but in a random direction. In this case we observe an even stronger effect of the social network on human mobility.

Influence of friends on an individual's mobility. We just observed a surprising increase of the effect of distant friends on an individual's mobility. However, so far we can only conclude that the relative correlation between one's travels and locations of her friends' homes increases with the distance traveled. Since friendships could have been created before or after the movement of a user, there are two possible explanations for this phenomena. If friendships are created first, then the explanation is that friendships influence where people travel. However, if movement precedes tie creation then it is traveling that influences/creates our social networks.

To distinguish between an existing friendship causing a user to move to a certain location and a movement to a certain place that then causes a formation of a new friendship, we obtained the Gowalla social network at two different time points t_1 and t_2 approximately three months apart. Comparing the network structure and check-ins between times t_1 and t_2 allows us to measure the degree of causality in each direction. We measure the effect of a user's existing friendship on her mobility by examining a set of check-ins C_a that occurred one day after time t_1 . Using check-ins in C_a we then measure what fraction of them occurred within the vicinity of friend's home, i.e., in radius $r = 25$ from one of the existing friends' homes. Since the network only contains edges created before time t_1 this quantifies the degree of influence of existing social network ties on future movement of a user. Similarly, we also examine the influence of mobility on creating new social network ties by examining a set of check-ins C_b on a day before time t_1 . We then count what fraction of check-ins in C_b led to creation of new friendships in the next three month period. We consider the mobility of user A to cause formation of a friendship (A, B) if at

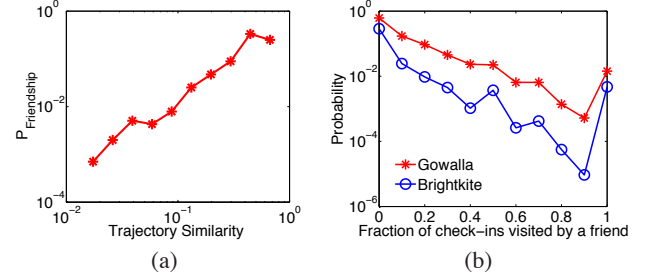


Figure 4: (a) Check-in similarity and probability of friendship. (b) Fraction of check-ins that were visited by a friend before a user checked-in at that same location.

least one of A 's check-ins in the set C_b occurred within radius r ($r = 25\text{km}$) of B 's home. Since the check-in preceded the formation of the social network tie we can consider this as the influence of movement on social network tie creation.

Performing the experiment on Gowalla we find that on average there is a 61% probability that a user will visit a home of an existing friend. However, the probability that a check-in will lead to a new formation of a new friendship is 24%. Basically, the influence of friendship on individual's mobility is about 2.5 times greater than the influence of mobility on creating friendships. We also performed a similar experiment using the cell phone data and consistently observed very similar behavior. We considered pairs of people living more than 100km apart and found that the probability of meeting an existing friend at a particular location is 70% higher than the probability of mobility creating new network ties.

Moving to where a friend has checked-in before. So far we examined the movement of A while we kept the locations of her friends B fixed, i.e., we only considered A going close to B 's home. Now we extend the setting and consider the case where both A and her friends B are simultaneously moving and we are interested in various aspects of the probability that they meet. Social check-in data gives us the exact location (e.g., the name of the coffee shop) a user has visited. We perform a similar set of experiments as before and examine the fraction of the user's check-ins have been visited by at least one of her friends prior to the user's first check-in to that specific place. We also devise a null model and make very similar observations as before. We find that the farther away a user travels the more likely that movement is influenced by a friend. The amount that friendship influences movement when traveling long distances (longer than 1,000km) is an order of magnitude higher than the influence when traveling short distances (shorter than 25km). Again, we also observe that for movement farther than 100km from home the probability of checking-in at the exact same location as a friend has checked-in in the past remains constant at around 10%. (We omit the plots due to brevity.)

Limits of using friendship for predicting mobility. Although we have shown the increasing influence of friends on one's mobility, only 9.6% of all check-ins in Gowalla and 4.1% of all check-ins in Brightkite were first visited by a friend and then by the user. To further explore this we now study the limits of using friendship for predicting human mobility. We observe strong correspondence between the trajectory similarity of users and probability of friendship. However, we also show that in general only a small fraction of users have a high overlap in check-ins with their friends.

Here, we explore the connection between trajectory similarity of a pair of users and the probability that they are connected in the social network. For each user we create a trajectory vector where the i -th component counts the number of times the user checked-in to location i . We then define trajectory similarity between a pair

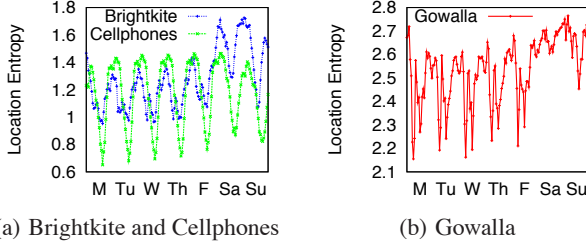


Figure 5: Average entropy of check-in locations over time. Notice periodicity and low location entropy in the mornings.

of users as the cosine similarity between corresponding trajectory vectors. Fig. 4(a) shows the probability of friendship as a function of trajectory similarity. We observe strong dependency between the two quantities, e.g., when a pair of users has more than 40% of check-ins in common then the friendship probability is above 0.3. We consider this a strong presence of social and geographical homophily [25].

Although we have shown a strong correlation between trajectory similarity and friendship, users have to have a certain amount of check-in overlap in order for the trajectory similarity to be meaningful. To investigate this we compute the fraction of each user’s check-in locations that were previously visited by a friend and then plot the distribution of this quantity in Fig. 4(b). We observe that the majority of the users have zero check-ins that were previously checked-in by a friend. In Gowalla, 84% of users have less than 20% of their check-ins that were previously visited by a friend, and 52% of the users have zero check-ins that were previously visited by a friend. This means that for about 50% of the users there is basically no information about their mobility that could be inferred from their social network.

Temporal and geographic periodicity of human movement. Our investigations of the effects of social network on human mobility so far have given mixed results. On one hand we found that social network influences long distance travel more than short distance travel, while on the other hand, we also observed that a relatively small fraction of user check-ins were previously checked-in by a friend. We now turn our attention to the non-social factors of human mobility. In particular, we explore periodicity (both temporal and geographical) of human mobility. Intuitively we expect that certain types of locations, such as home and work, are visited regularly, and often during the same times of the day.

One way to quantify the periodicity is to measure the fraction of user check-ins that are visits to previously already visited locations. We observe that 53% of all check-ins in Brightkite (31% in Gowalla) have been previously visited by the same user. This means that if a user checks-in into a place for the first time, there is 53% chance she might return and check-in again. On the other hand, the effect of the social network is about 5 times smaller, i.e., as we have shown before there is only a 10% chance that a user will check-in to a place where a friend has checked-in before.

Last, we also explore the connection between geographic and temporal periodicity. Figure 5 plots the average Shannon entropy of check-in locations for each hour of the week. The lower the entropy, the lower the variability of check-ins during that time period. We observe the periodicity on both the 24 hour and the 7 day scale. For all days, the early morning hours have the lowest location entropy (i.e., most people are at home). Location entropy increases when people are commuting during the rush hour and in the evening when they might be out socializing. During the work week entropy is lower compared to that of the weekend (at least for the Brightkite and Gowalla data). This suggests that people are

commuting to and from work at roughly the same time during the work week, as opposed to the weekend when peoples’ travel and schedules are less predictable.

4. MODEL OF HUMAN MOBILITY

In the following section we develop a model of human mobility that can accurately predict future movements of an individual. Our results so far have given strong evidence of geographic (returning to the same places) and temporal (traveling at regular times of the day) periodicity as well as the increasing relative effect of the social network structure on an individual’s mobility. We proceed by formulating a coherent model that incorporates the three essential ingredients of human mobility: temporally and geographically periodic movement with the social network structure. First, we present the *Periodic Mobility Model (PMM)* that we later extend to *Periodic & Social Mobility Model (PSMM)* to also account for the mobility due to the social network structure.

Periodic Mobility Model (PMM). We build on the intuition that the majority of human movement is based on periodic movement between a small set of latent states (locations). For simplicity we introduce the model using just two latent states (even though our model can handle an arbitrary number of them). One can simply think of the two latent states as “work” and “home”. Depending on the time of the day, an individual’s movements will either be centered around home, work, or somewhere in between the two locations as they “commute” in between them.¹ Fig. 6 illustrates the model: (a) for each user we infer the centers of geographic locations of two latent states and model them with a gaussian distribution; (b) we model the probability of whether the user is in the home/work state as a function of the time of day. Fig. 6(a) shows a user in San Francisco that lives in the Mission (red crosses) and works in the Financial district (blue circles). For each check-in, we infer whether it was generated while in the home or work state, and the squares denote the geographic center of home/work check-ins. Fig. 6(b) illustrates the temporal model of how the user transitions between the two states with the distance from the center of the circle being proportional to the probability that a user resides in that state (user stays at “home” overnight (blue) and at “work” (red) during the day).

We can also think of our model in a generative way. A user determines the location of her check-in in the following way: given the time of the day, the user first samples whether she is in the “home” or “work” state (Fig. 6(b)). Then, depending on the state the user samples the geographic position of her check-in from the appropriate home/work location distribution (Fig. 6(a)).

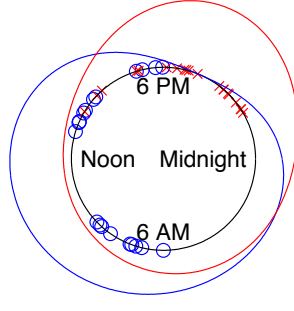
We now proceed by formally introducing the *PMM* model. Let t be the current time of the day and let $x_u(t)$ denote the geographic position of user u at time t . Let $c_u(t)$ be the “state” at time t , where $c_u(t) = H$ denotes that the user is in the “home” state at time t , and $c_u(t) = W$ indicates the user is in the “work” state. The spatial location of user’s check-ins is then governed by the appropriate home/work check-in location distribution $P[x(t) = x|c_u(t)]$. The probability distribution over the locations of a user at time t is simply the mixture of “home” and “work” location distributions where mixing coefficients are governed by the temporal model (belief over the current state of the user):

$$P[x(t) = x] = P[x_u(t) = x|c_u(t) = H] \cdot P[c_u(t) = H] + P[x_u(t) = x|c_u(t) = W] \cdot P[c_u(t) = W].$$

¹Note that geographic locations of latent states are unknown and we have to infer them. Also, the latent states simply correspond to latent positions around which most check-ins are centered. We name the states “home” and “work” for easier exposition.



(a) Spatial model



(b) Temporal model

Figure 6: (a) Check-ins of a user in San Francisco: geographic distribution of check-ins when in home/work state. (b) Temporal model: distance to the red/blue line from the center is proportional the prob. of user being in home/work state.

This means that at any point in time a user is either in “home” or “work” state and $P[c_u(t)]$ models the probability distribution over the state of the user over time. We then maintain the distribution over possible check-in locations independently for both states.

Temporal component of the PMM model. We model $P[c_u(t)]$, the probability distribution over the state of the user, with a truncated Gaussian distribution parameterized by the time of the day:

$$N_H(t) = \frac{P_{c_H}}{\sqrt{2\pi\sigma_H^2}} \exp\left[-\left(\frac{\pi}{12}\right)^2 \frac{(t - \tau_H)^2}{2\sigma_H^2}\right]$$

$$N_W(t) = \frac{P_{c_W}}{\sqrt{2\pi\sigma_W^2}} \exp\left[-\left(\frac{\pi}{12}\right)^2 \frac{(t - \tau_W)^2}{2\sigma_W^2}\right]$$

and then

$$P[c_u(t) = H] = \frac{N_H(t)}{N_H(t) + N_W(t)}$$

$$P[c_u(t) = W] = \frac{N_W(t)}{N_H(t) + N_W(t)}$$

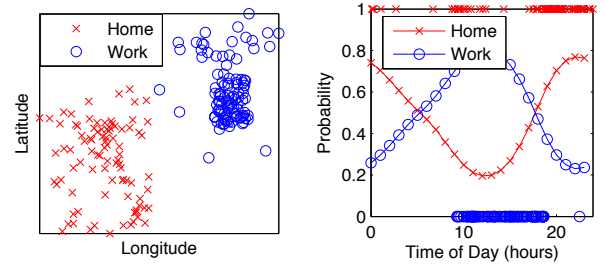
where τ_H is the average time of the day when a user is in the “home” state, σ_H is the variance in time of day, and P_{c_H} is the time-independent probability that any given check-in was generated by the “home” state. It should be noted that special consideration is needed when calculating τ_H because we are treating the time of day as cyclical, i.e., 11:59pm and 12:01am are only two minutes apart. Therefore, τ_H and τ_W are calculated in the same manner as an average of angles about a circle. For example, Figure 7(b) shows the probability distribution over the state (blue:home, red:work) of the user as a function of the time of the day.

Spatial component of the PMM model. Many previous papers [12, 2] have used normal distributions to model human movement around a particular point. We adopt this by modeling the movement when a user is in the home/work state using a 2-dimensional time-independent Gaussian distribution:

$$P[x_u(t) = x_i | c_u(t)] = \begin{cases} \sim \mathcal{N}(\mu_H, \Sigma_H) & \text{if } c_u(t) = H \\ \sim \mathcal{N}(\mu_W, \Sigma_W) & \text{if } c_u(t) = W \end{cases}$$

where Σ_H, Σ_W are the “home”, “work” check-in position covariance matrices. μ_H and μ_W are the means of user’s check-in locations when she is in home and work state, respectively.

Our model is effectively a two-state mixture of Gaussians with a time-dependent state prior. This means that our model classifies each of the user’s check-ins as either being generated by the “home” or “work” state. The temporal part of the model governs the transition between home/work states and then depending on



(a) Location of user check-ins

(b) Home/work state

Figure 7: Periodic Mobility Model. (a) Check-in locations generated by home/work state. (b) State distribution over time.

the state geographic location of the check-in is generated the time-varying mixture of two time-invariant 2-dimensional Gaussian distributions. Figure 8 shows distribution of user location over time.

Periodic & Social Mobility Model (PSMM). Next, we extend the *Periodic Mobility Model* with social network-driven mobility. We refer to the new model as *Periodic & Social Mobility Model (PSMM)*. To include the social network information to the model, we introduce another check-in classification $z_u(t)$, where $z_u(t) = 1$ implies the check-in is social (non-periodic) and $z_u(t) = 0$ implies that it is periodic. The *PSMM* mobility model then becomes:

$$P_u[x(t) = x] = P[x(t) = x | z_u(t) = 1] \cdot P[z_u(t) = 1] + P[x(t) = x | z_u(t) = 0] \cdot P[z_u(t) = 0]$$

where $P[x(t) = x | z_u(t) = 0]$ is the *Periodic Mobility Model*.

Given that user u makes a social check-in (i.e., social network influenced check-in), the probability that u will check-in at a certain place x_i is determined by two factors: how long since a friend, say w , has checked in, and the distance of w ’s check-in to x_i . We model this as follows:

$$P[x_u(t) = x_i | z(t) = 1] \sim \sum_{(t_j, x_j) \in J_u} |t_j - t|^{-\alpha} \cdot ||x_i - x_j||^{-\beta}$$

where J_u is the set of check-ins by user u ’s friends made on the same day. t_j denotes the time and x_j is the location of the j -th checkin by u ’s friends. The particular power-law parametric forms are motivated by the following observations: the probability user u checking-in Δt time units after w has checked-in decays as power-law (Fig. 9(a)) and Fig. 9(b) shows that the distribution of the distance between two friends at the time of a cell phone call also decays as power law. From the modeling perspective this means that if user u performs a social check-in then it will more likely be close in space and in time to one of her friend’s check-ins.

Fitting PMM and PSMM models. First, we describe how to fit the parameters of the *PMM* independent of the social network and then consider the fitting of social check-ins. The parameters of the model are fitted using Expectation-Maximization (EM). To begin, each check-in is randomly labeled as either a “home” or “work” check-in. At each iteration, first the model parameters μ, Σ, τ , and σ are fitted by the current labeling using maximum likelihood estimation. This parameter fitting is known as the “E-step.” All of these parameters MLE’s have closed-form solutions, making this step very efficient. Once the model parameters have been fitted, the check-ins are reassigned to the state (home/work) according to the new model density function. This labeling reassignment is known as the “M-step”. With this new check-in labeling, the “E-step” is once again performed, and iterations continue until convergence. Since EM is known to only converge to local optima, we re-run the algorithm using several different random initial assignment of check-in labels and take the fit with the highest likelihood.

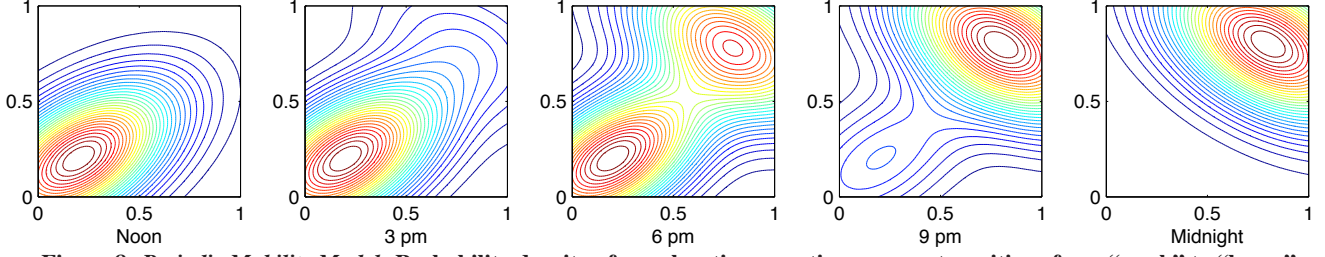


Figure 8: *Periodic Mobility Model*: Probability density of user location over time as user transitions from “work” to “home”.

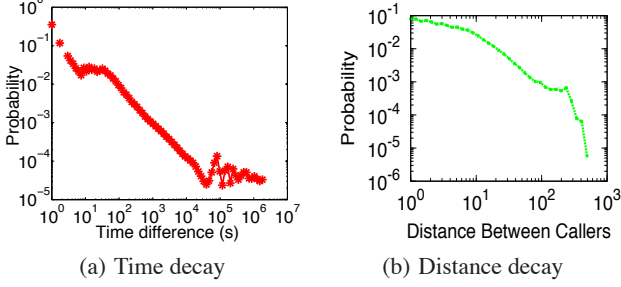


Figure 9: **Temporal and spatial properties of check-ins:** (a) Time since last friend has checked-in to the same place. (b) Distance between two users at the time of a call.

PMM has 18 parameters: 4 parameters for temporal model (mean and variance of time when the home/work state occurs), 12 for the spatial model (two 2-d means and corresponding covariance matrices) and 2 parameters for the social model (time and distance decay). In order to prevent overfitting we use several regularizations: we impose a minimum singular value of 10^{-7} on Σ (usually, this value is naturally several orders of magnitude larger), and we bound the temporal variance σ to be above 10^{-4} .

We fit *PSMM* as follows. We train the original *PMM* model where we allow a check-in to be classified as “home”, “work” or “outlier” (i.e., social) check-in. Thus we assume that check-ins that are not fit well with the periodic model are the result of social activity. We then fit the social model to these “outlier” check-ins. Overall, 10%-30% of all check-ins are classified as social check-ins. The relabeling of the social check-ins and the fitting of the model parameters α and β is done for each user through EM.

5. EXPERIMENTAL EVALUATION

In the following section, we evaluate the proposed *PMM* and *PSMM* models on all three datasets. In particular, we are interested in the predictive performance of the models, i.e., given the time of the day we aim to as accurately as possible predict the geographic location of the user. We consider three evaluation metrics and three strong baseline location prediction methods for comparison.

Evaluation metrics. To compare different mobility models, we use the following evaluation metrics. First, we consider the average log-likelihood of the check-ins in the unseen test set. This measures how well the test set fits the model. The second metric we consider is predictive accuracy, i.e., given the time of day of a check-in in the test set, how accurately can each model predict the exact location of the check-in. For example, accuracy of 0.5 means that 50% of the time the model correctly predicts the exact check-in location (out of all known locations). Accuracy is a very harsh metric. For example, if a model would always predict the true location with exactly 1 meter error, the accuracy of such model would be zero. The third performance metric we consider is the Expected Distance Error, which can be considered a soft version of accuracy in that

it does not insist on predicting the exact location, and it takes into account the spatial proximity of predictions to actual check-ins. We define Expected Distance Error as follows. For a given check-in in the test set, we measure expected distance between the check-in and a check-in generated from the model’s probability distribution:

$$\begin{aligned} d(P) &= \frac{1}{|C_{te}|} \sum_{c \in C_{te}} E_{P(x)} [\|x_c - x\|] \\ &= \frac{1}{|C_{te}|} \sum_{c \in C_{te}} \int_x \|x_c - x\| \cdot P(x) dx. \end{aligned}$$

Since it is computationally infeasible to calculate the exact expected distance error across each point on the entire globe, we compute the error over all locations ever checked-in by the user.

One issue with this metric is that it does not consider the typical distance a user travels. Thus, using the expected error distance metric to compare the performance of a model across two different users is uninformative if the two users travel on different scales. For example, if user typically travels 1,000km then an average distance error of 10km is much more acceptable than the same error for someone who typically travels only 20km. To correct for this we normalize the error by the radius of gyration [12], which is the average distance a user travels on a particular day. For the expected distance error of each user, we divide it by the user’s radius of gyration for that particular day of the week, creating the *relative* expected distance error.

Baseline models. We also consider three non-trivial baseline models for comparison. The first baseline, which we refer to as the *Most Frequented Location Model (MF)* assigns the probability of a user checking-in at a location x_c during a given hour of the day h_c as the fraction of previous check-ins during that hour h_c that were at location x_c . More precisely, let C_u be the set of all check-ins of user u , then the Most Frequented Location Model is

$$P_M[x_u(t) = x | t \in h] = \frac{|\{c | c \in C_u, x_c = x, h_c = h\}|}{|\{c | c \in C_u, h_c = h\}|}.$$

Despite its simplicity, this model is a very strong baseline. For every hour of the day the *MF* predicts the most likely (most often checked-in) location of a particular user. The model is also very intuitive as it imitates the following reasoning: Assume that you are asked to guess the location of a friend at 2pm on a Wednesday. If you knew their place of work, you would be able to pin point them with a high level of certainty. Similarly, if you knew the location of your friend’s most frequented happy hour spot, locating them at 6pm on a Friday would be easy as well. In fact, if a user is perfectly periodic, then as $|C_u| \rightarrow \infty$, P_M will converge to the true underlying model. What this model lacks (and what our model corrects) is that it does not consider distance or the spatial proximity of locations.

The second baseline, the *Gaussian Model (G)*, has been proposed by Gonzales et al. [12]. It models human movement as a stochastic process centered around a single point. For a given day of the week,

let μ_x be the mean location of check-ins that occurred on that day of the week, and let Σ_x be the covariance matrix:

$$P_M[x_u(t) = x] \sim \mathcal{N}(\mu_x, \Sigma_x).$$

Note that this model is static in time and captures the scale of a user’s movements more than anything else.

Lastly, we also consider a model that could potentially have very good performance. The *RW* model always predicts the next location of a user to be the location of her last known check-in location.

Experimental setup. In all our experiments we only consider users that had at least 10 check-ins on each day of the week. We put 80% of their check-ins in the training set and the remaining 20% in the test set. This means that we evaluate the models on 6,233 Brightkite, 10,997 Gowalla, and 853,812 Cellphone users. For each user we build 7 models, one for each day of the week. Overall, this means that at minimum we use only 8 check-ins (on a particular day) to build a model and remaining 2 to test the model.

Predicting mobility. First, we compare the performance of the periodic *PMM* model with the performance of the baselines. For brevity we only discuss results for Brightkite and Cellphones.

Figure 10(a) shows the accuracy for all four models for Brightkite. The Gaussian model (*G*), proposed by Gonzalez [12], performs worst with an average accuracy of 19.7%, predicting last known location of the user (*RW*) gives 22.7% accuracy, while the most frequent location model (*MF*) scores 32.5%. Our model gives an 83.1% improvement over the Gaussian model and 11.0% relative improvement over the *MF* model. Furthermore, *PMM* can determine the *exact* location of a check-in up to 40% of the time. Considering how much noise is in the data and the large number of different locations a user visits, we consider this a remarkable result. Interestingly we also observe that all models exhibit better performance on weekdays while predicting user location is harder on the weekends.

Similarly, Fig. 10(b) shows the accuracy for the Cellphones. We make similar observations. Again, *PMM* does 110.9% and 44.3% better than the *G* and *RW* models, respectively, with overall accuracy of 42%. However, the difference between *PMM* and *MF* is smaller. This is due to the coarser spatial resolution of the cellphone data: mobility is captured in coarser detail and thus less location “smoothing” is needed.

Figures 10(c) and 10(d) give results for test data log-likelihood evaluation metric. The Gaussian model (*G*) performs far the worst, with the *PMM* outperforming the two baselines on both datasets. (*RW* is not included because it has no density function.)

Last, Figures 10(e) and 10(f) display the Relative Expected Distance Error for the four models. Average relative distance error of *PMM* is 2.9% on Brightkite and 0.27% on cellphones. Overall, this is 55.6% better than Gaussian model, 28.9% better than last location model (*RW*) and 27.4% better than most-frequent location model (*MF*). In absolute terms, the average absolute location error of *PMM* is about 10km on Cellphones and 90km on Brightkite.

Predictive performance of the social model. For both average log-likelihood and the relative expected distance error, *PSMM* provides considerable improvement over *PMM* for both Brightkite and Cellphones — about a 10% relative improvement for the relative expected distance error and about a 25% relative improvement for log-likelihood. We expect that with denser data sets (more check-in data), the occurrence of two friends checking-in at the exact same place will become more frequent, and then *PSMM* will yield even higher improvement over the periodic model.

Similarity of daily mobility patterns. We also quantify how similar are mobility patterns during different days of the week. We establish this using the following model “similarity” metric: Let

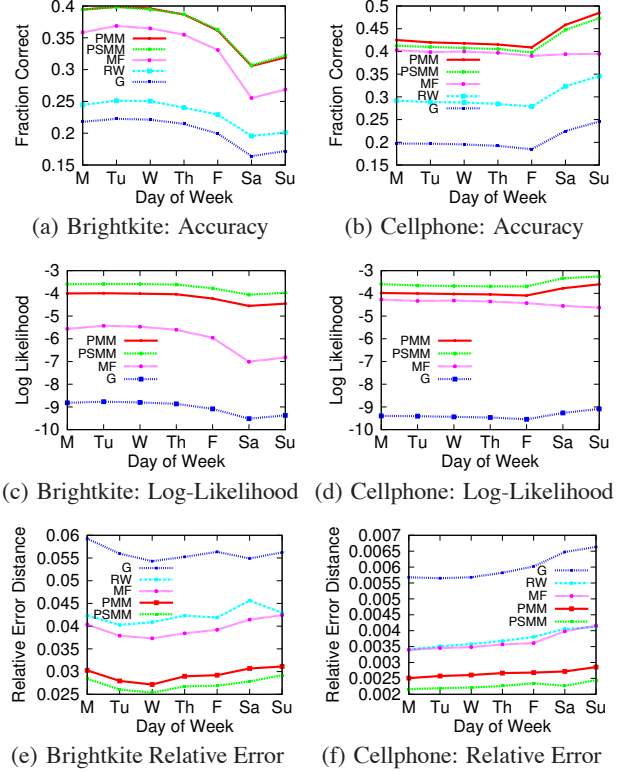


Figure 10: Performance of the Periodic Mobility Model (*PMM*) and the Periodic & Social Mobility Model (*PSMM*), compared to three baseline models. (a,b) Accuracy of check-in location prediction; (c,d) Log-likelihood of check-ins in the test set; (e,f) Expected Distance Error of predicted check-in location.

C_1 and C_2 represent two different sets of check-ins, and let P_1 and P_2 be the *PMM* fitted to these two sets. Then we define

$$S(P_1, P_2) = \frac{\sum_{c \in C_1 \cup C_2} P_1(c) \cdot P_2(c)}{\sqrt{\sum_{c \in C_1 \cup C_2} P_1(c)^2} \sqrt{\sum_{c \in C_1 \cup C_2} P_2(c)^2}}.$$

S measure is analogous to covariance sampled at a subset of points. S measure can compare any two *PMM* models, whether they are models for different days of the week for one user, or the same day of the week for two different users. $S(P_1, P_2) = 1$ if $P_1 = P_2$.

Figure 11 represents the average S between the models for each day of the week for high volume users of (a) Brightkite and (b) Cellphones. The work days show strong correlation with each other, but less correlation with the weekend days. Similarly, the weekend days (Saturday and Sunday) show a stronger correlation with each other than they do with the other days. This implies that even in the absence of a commute to and from work, there is still periodic structure to human movement.

Number of latent states. So far we only considered the model with two latent states (“home” and “work”) but it seems very natural to consider richer models with additional latent states. Indeed, when 3, 4, etc. latent states are used (each latent state with its own spacial Gaussian and temporal periodic prior) the accuracy of the model increases. For each latent state that is added, however, the returns are diminishing. Even the performance gained by adding the third latent state to the model is minor compared to the difference between the two state model and a single state model. For example, the average accuracy across all days of the week for Brightkite is 19.7% using a single state model, whereas the two state model

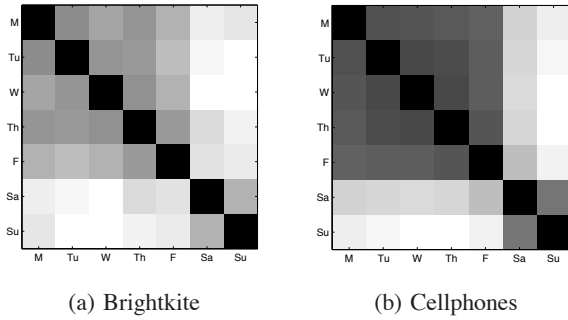


Figure 11: Average cross-similarity between a user's periodic models constructed for each day of the week. The darker the square, the stronger the similarity. Note that the work days are more similar to each other than to the weekend days.

has an accuracy of 36.1%, which is a 83.1% relative improvement. Using a three state model only increases performance by another 3.3%. Thus we conclude that the two state “home” and “work” model is rich enough to capture the human mobility.

6. CONCLUSION

We investigated patterns of human mobility on three large but very different datasets: two sets of location-based social network check-in data spanning the whole planet and also cellphone location data. Even though location-based social networking services are very different from cell phone tower location data, we found many common patterns of human mobility across the datasets. Most surprisingly, we found that humans experience a combination of strong short range spatially and temporally periodic movement that is not impacted by the social network structure, while long-distance travel is more influenced by the social network ties.

We also developed a model of human mobility dynamics. The model combines the periodic day-to-day movement patterns with the social movement effects coming from the friendship network. Our model reliably captures and predicts human mobility patterns and outperforms current mobility models by a factor of two.

Acknowledgements

Research was in-part supported by NSF CNS-1010921, NSF IIS-1016909, Microsoft Faculty Fellowship, Albert Yu & Mary Bechmann Foundation, Intel, IBM, Lightspeed and Yahoo. Eunjoon Cho was supported by the Samsung Scholarship.

7. REFERENCES

- [1] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: Improving geographical prediction with social and spatial proximity. In *WWW '10*, pages 61–70, 2010.
- [2] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 2006.
- [3] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 2009.
- [4] N. Couldry and A. McCarthy. *Mediaspace: Place, Scale and Culture in a Media Age (Comedia)*. Routledge, Nov. 2003.
- [5] D. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *PNAS*, 2010.
- [6] J. Cranshaw, E. Toch, J. I. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *UbiComp*, pages 119–128, 2010.
- [7] N. Eagle and A. Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Soc.*, 2009.
- [8] N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *PNAS*, 2009.
- [9] S. Eubank, H. Guclu, V. S. Anil Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 2004.
- [10] E. Glaeser and M. Kahn. Sprawl and urban growth. *Handbook of regional and urban economics*, 2004.
- [11] J. Goldenberg and M. Levy. Distance is not dead: Social interaction and geographical distance in the internet era. *Arxiv*, abs/0906.3202, 2009.
- [12] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
- [13] Q. Hao, R. Cai, C. Wang, R. Xiao, J.-M. Yang, Y. Pang, and L. Zhang. Equip tourists with knowledge mined from travelogues. In *WWW '10*, pages 401–410, 2010.
- [14] T. J. Hatton and J. G. Williamson. Global Migration and the World Economy: Two Centuries of Policy and Performance. The MIT Press, 2008.
- [15] L. Humphreys. Mobile social networks and social practice: A case study of dodgeball. *Journal of Computer-Mediated Communication*, 2008.
- [16] B. Jiang, J. Yin, and S. Zhao. Characterizing the human mobility pattern in a large street network. *Physical Review E*, 80(2):021136, Aug. 2009.
- [17] A. Kapoor, N. Eagle, and E. Horvitz. People, quakes, and communications: Inferences from call dynamics about a seismic event and its influences on a population. In *AAAI Spring Symposium*, 2010.
- [18] M. Kim, D. Kotz, and S. Kim. Extracting a mobility model from real user traces. *INFOCOM '06*, 2006.
- [19] R. Lambiotte, V. Blondel, C. Deckerchov, E. Huens, C. Prieur, Z. Smoreda, and P. Vandooren. Geographical dispersal of mobile communication networks. *Physica A*, 387(21):5317–5325, Sept. 2008.
- [20] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong. SLAW: A new mobility model for human walks. In *INFOCOM '09*, pages 855–863, 2009.
- [21] T. Leighton. Improving performance on the internet. *Queue*, 6:20–29, 2008.
- [22] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *WWW '08*, 2008.
- [23] Z. Li, B. Ding, J. Han, R. Kays, and P. Nye. Mining periodic behaviors for moving objects. *KDD '10*, 2010.
- [24] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *PNAS*, 2005.
- [25] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.
- [26] D. Mok, B. Wellman, and J. Carrasco. Does distance matter in the age of the internet? *Urban Studies*, 47(13), 2010.
- [27] A. J. Nicholson and B. D. Noble. Breadcrumbs: Forecasting mobile connectivity. In *MOBICOM '08*, pages 46–57, 2008.
- [28] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. In *ICWSM '11*, 2011.
- [29] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. In *ICWSM '11*, 2011.
- [30] C. Song, Z. Qu, N. Blumm, and A. Barabási. Limits of predictability in human mobility. *Science*, 2010.
- [31] B. Wellman and K. Hampton. Living networked on and off line. *Contemporary Sociology*, 28(6):648–654, Nov. 1999.
- [32] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative location and activity recommendations with GPS history data. In *WWW '10*, pages 1029–1038, 2010.
- [33] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from GPS trajectories. In *WWW '09*, pages 791–800, 2009.