Bridging the Gap Between Physical Location and Online Social Networks

Justin Cranshaw, Eran Toch,
Jason Hong, Aniket Kittur, Norman Sadeh

March, 2010
CMU-ISR-10-107

Institute for Software Research
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

An expanded version of this work appeared in the Proceedings of the 12th ACM International Conference on Ubiquitous Computing, Copenhagen, Denmark, September 2010.

This work is supported by NSF Cyber Trust grant CNS-0627513, by CyLab at Carnegie Mellon under grant DAAD19-02-1-0389 from the Army Research Office, and by the Fundação para a Ciência e a Tecnologia through the Carnegie Mellon Portugal Program. Additional support has been provided by Microsoft through the Carnegie Mellon Center for Computational Thinking, and by grants from France Telecom, Nokia and Google.

Key Words: Location sensing, Location Tracking, Social network analysis, Social computing, Human-computer interaction

Abstract

This paper examines the location traces of 489 users of a location sharing social network for relationships between the users' mobility patterns and structural properties of their underlying social network. We introduce a novel set of location-based features for analyzing the social context of a geographic region, including location entropy, which measures the diversity of unique visitors of a location. Using these features, we provide a model for predicting friendship between two users by analyzing their location trails. Our model achieves significant gains over simpler models based only on direct properties of the co-location histories, such as the number of co-locations. We also show a positive relationship between the entropy of the locations the user visits and the number of social ties that user has in the network. We discuss how the offline mobility of users can have implications for both researchers and designers of online social networks.

1 Introduction

Although voices in the media and academia often make distinctions between *online* social networks and *offline* social networks, until recently it has been extremely difficult to rigorously address questions comparing these two worlds. This has led to conflicting results when researchers have attempted to relate online and offline behavior. For example, in a recent article Deresiewicz argues that online social networks are contributing to the isolation of people in the physical world [2], while a recent Pew Internet and American Life report argues that online social networks have a positive impact on social relations in the physical world [9]. The current lack of methodology for analyzing the distinctions between online and offline social networks can explain, in part, this type of open ended debate.

At the same time, the growing ubiquity of location-enabled "smartphones" blurs the distinction between online and offline social networks. This is most apparent in emerging mobile social networks such as Foursquare and Gowalla which have created new means for online interaction based entirely on the physical location of their users. Furthermore, smart devices also make it possible to study peoples' offline behavior by continuously tracking their whereabouts. As a consequence, many questions of human behavior which in the past were difficult to answer, will soon become easier to analyze.

One of the challenging problems in this space is inferring properties of the social behavior of users from their location trails. Some promising research in this area can be seen in papers by Eagle et al. [3, 4] and Li et al. [11] who develop measures of user similarity based on mobility and use this to infer the social structures of the users. This task is particularly challenging since co-location of two users, loosely defined as being in the same place at the same time, does not provide enough evidence to reliably establish a relationship between them, especially in urban environments, where co-location among strangers is frequent [12]. Furthermore, in realistic conditions, location tracking is inherently partial and inexact, making this kind of inference difficult on a large scale.

To meet these challenges we introduce a set of features that shed light on the social context of the locations that users visit. We evaluate these features on two tasks: predicting whether two co-located users are friends on Facebook, and predicting the number of friends a user has in the social network. Additionally, we examine the relative importance of the predictors used, and we show that looking deeper into characteristics of the locations the users visit can significantly improve performance on these tasks.

Being able to rigorously address these questions requires a special experimental framework capable of observing both the offline social behavior and the online social structure of the users. To meet this end, we use Locaccino, a location-sharing application based on Facebook's social network [14]. Locaccino allows users to share their location with their Facebook friends subject to robust privacy preferences. Our results are based on an analysis of the location trails of 489 participants who were tracked using GPS and WiFi positioning technologies installed on the their mobile phones and/or laptop computers.

In this work we introduce and evaluate a set of contextual features of human location trail data for inferring two social aspects of the users: the existence of an online social network link between two users, and the number of friends a user has. We show that by analyzing characteristics of the locations the users visit, and by studying the patterns of an individual user's mobility, we can gain

valuable context into the users social world.

This work makes the following primary research contributions:

1. We establish a model of friendship in an online social network based on contextual features of user co-location.

2. We identify positive relationships between the mobility patterns of a user and the number of online friends the user has.

3. We show that diversity measurements of a location, such as the entropy of the distribution of unique visitors there, can be used to analyze the context of the social interactions at that location.

2 Related Work

Several promising results demonstrate the potential of using ubiquitous mobile technologies to study human social behavior. In a series of papers, González et al. observed a large group of mobile phone users over six months, showing that phone users' mobility patterns have a high degree of spatial and temporal regularity [7]. They then used this insight to developed statistical models of user mobility patterns. Eagle and Pentland used eigenvalue decomposition to study routine behaviors of mobile phone users [3]. They showed that the inferred principal components of participant behaviors discovered by their decomposition can be used to build a similarity measure between users. Furthermore, they showed that this similarity measure can be used to successfully infer familiarity. Li et al. also used location histories to derive a user similarity measure [11]. Their similarity measure is derived from a hierarchical modeling of the users' location histories that takes into account both movements on a micro scale, say from building to building, and movements on a macro scale, say from city to city. Miklas et al. studied the network of interactions of mobile phone participants in relation to their social network [12]. Although the primary focus of their work was on applications that exploit social interactions such as routing in delay tolerant networks, they also found several interesting descriptive results about social interactions, such as the distribution of participant interactions with strangers versus interactions with friends.

Eagle et al. analyzed a set of features of mobility data to study the social structure of the participants [4]. They examined features such as the proximity of the users at work, proximity on a Saturday night, whether there was phone communication between them, and the number of unique locations where they were together as predictors of whether there was a relationship between the two users. They then conducted a regression analysis using self report data for the actual user relationships to study what factors contribute most to friendship. Their analysis showed that phone communication was by far the most significant predictor of friendship, followed by the number of unique location, and proximity on a Saturday night.

We build on this foundation and expand it several ways. First, we compare physical social interactions with an existing online social network rather than self reported social ties. Not only does this bypass any potential biases introduced by self report data, this type of analysis also allows our

work to contribute new applications to online social networks, such as location-based friend recommendation and categorization systems, and location recommendation systems. Second, we do not record the existence of cellular phone communication between the users. Rather, our methodology is based only on knowing the users' locations. Finally, we expand the existing methodology for analyzing location data, by introducing new tools for enhancing the understanding of the context of human location observations by looking at global properties of the location where the observation occurred, such as the entropy of the distribution of users that visit the location. We show that using these new location-based features, we can construct a classifier for predicting social ties that outperforms one that is based on features similar to the proximity-based features used used by Eagle et al. [4].

Unlike the bluetooth handshake method for inferring interactions used by Eagle et al. [4] which requires communication between the phones of the participants in order to establish proximity, our method records the location of the users using standard GPS and WiFi geo-positioning, similar to Li et. al [11]. We then infer by proximity, rather than explicitly observe via bluetooth handshake, the social interactions between the users. This method is realistic and highly scalable, making it relevant for researchers and practitioners wishing to study user location traces on a large scale. Most importantly, although inferring social interaction in this way can produce noisy data, we show how a sophisticated analysis of the context of the observed proximity can compensate for data limitations.

The methodology we present in this paper also offers new tools that can be used in future research on the impact of the Internet on social relations. Current research in this field is is based mostly on qualitative findings and surveys. For example, Barry Wellman et al. studied the impact of Internet on neighborhoods and families [15], Kraut et al. studied the effects of the Internet on the well being of users [10], and Ellison et al. studied social capital of Facebook [5]. While our current paper does not aim to contribute directly to any of these questions, our work provides another dimension to address these difficult questions in future work.

3 Method

We observed users through continuous tracking of their location using laptop computers and smart phones. Additionally we observed the existence of Facebook friendships between pairs of users. In this section, we describe in detail the technical and experimental framework, and the collected data.

3.1 Locaccino

Locaccino [14] is a Web-application developed by the Mobile Commerce Lab at Carnegie Mellon University that allows a user to share her current location with other Locaccino users through her Facebook social network subject to user-controllable privacy rule specifications[1]. From the user's perspective, there are two components of Locaccino: the *web application*, which allows users to

[1] www.locaccino.org

query their friends' locations and set up and review privacy rules, and the *locator software*, which runs on laptops and mobile phones (Symbian OS and Android) and updates the user location every 10 minutes.

Users run the client locator software in the background of their laptops or smart phones, which uses a combination of GPS (if available), WiFI, and IP geolocation to ascertain location coordinates of the user. Each method has differing levels of accuracy. Locations ascertained via GPS are typically accurate to within 10 to 15 meters. Locations ascertained through a WiFi lookup service like that provided by Skyhook Wireless[2] are typically accurate to within 10 to 20 meters. Locations ascertained via IP geolocation are typically at the city or neighborhood level of granularity. These location observations, which consist of a time-stamp together with latitude and longitude values, are sent to the Locaccino server by the client software in 10 minute intervals.

3.2 Recruitment, Demographics and Data Collection

The 489 users discussed in this work were each active users of Locaccino for periods ranging from 7 days to several months (mean of 74 days, median of 38 days). The participants started using Locaccino at different times and for different reasons. 285 of the users were recruited as part of 3 different studies from the campus population using fliers and posting on the university's electronic message boards. The rest of the users were either invited by study participants through a built-in invite mechanism, or they found Locaccino through research publications, online press, or other means. Figure 1 shows a plot of the number of unique users being tracked for each day of the period we study in this work. All users of Locaccino, regardless of how they were recruited, gave informed consent to participate in the study prior to registering an account on the system.

Although we recognize this might limit the generality of our findings, to enforce some control over the data we ignore all observations outside of the Pittsburgh metropolitan region (where Locaccino was first deployed). This allows us to study the users in a closed "ecosystem" and it frees the data from any bias that might result from an uneven density of observations across geographic regions. In total, over 3 million location observations were collected for this work, with nearly 2 million of these falling in the Pittsburgh region. Additionally, we ignore all location observations that were obtained by IP geolocation. Assuming each data point represents a 5 minute interval, this is over 20 years of cumulative human observational data.

A large percentage of the observations were collected from the laptop locator software (93.7%). This imposes several limitations on the data analysis. For one, people are not as mobile with laptops, which are often only powered on in stationary locations, as they are with cellular devices, which often remain powered-on and near that person at all times. Furthermore, laptops offer a much more sporadic approximation of a person's location than cellular devices do. Laptops are sometimes powered on for hours at a time while the user is in fact not near the laptop (for instance at home, or at the office). Although, this adds a significant element of noise to the data that is difficult to quantify, the data is nevertheless realistic, as it represents a real world deployment of a location sharing system. Furthermore, we feel the limitations of the data are testament to the strength of our methods, as we are able to find significant and strong results using data generated

[2] www.skyhookwireless.com

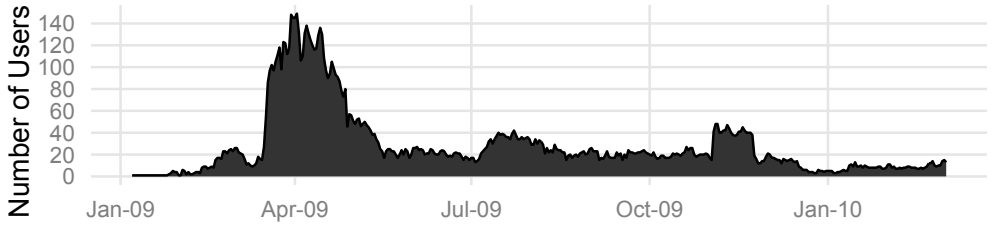

Figure 1: A plot of the number locatable study users in the Pittsburgh metropolitan region for each day of the study period. Peaks in Apr-09, Jul-09, and Dec-09 indicate controlled studies.

Graph Structural Properties	$\mathcal{S}$	$\mathcal{C}$	$\mathcal{S} \cap \mathcal{C}$
Number of vertices	397	397	397
Number of isolated vertices	15	120	206
Number of edges	**1063**	**3636**	**307**
Num connected components	106	108	234
Largest component size	315	275	67
Density	0.014	0.046	0.004
Connectedness	0.63	0.48	0.04
Degree centralization	0.06	0.22	0.03
Eigenvector centralization	0.42	0.21	0.50

Table 1: Structural properties of the networks analyzed in this work.

in this realistic and highly scalable manner.

3.3 Co-location

We divide the latitude and longitude space into discrete
0.0002×0.0002 latitude/longitude grids (approximately 30 meters $\times$ 30 meters) and the time coordinate into whole 10 minute intervals. In this way, a co-location of two users is defined as an observation of the users within the same 0.0002×0.0002 location grid within the same discrete 10 minute interval. The particular choice of discretization was based on practical considerations balancing the accuracy of the location sensing technology with the noise associated with larger discretization windows. Although such a discretization adds some noise when trying to infer co-locations, when examining the entire history of co-locations between pairs of users, this noise is marginalized. Unless otherwise stated, when we refer to a *location* or *location observation* or *co-location observation* in this paper we assume the location and time coordinates are subject to this discretization.

3.4 Network Data

In this work we primarily focus on network data induced from Locaccino user observations. In particular, we compare the network formed by co-location of system users, with the underlying Facebook social network. The three networks that we consider are defined below:

The Social Network: We denote the underlying *Facebook social network* of Locaccino users by $\mathcal{S}$. There is an edge between vertices $u_1, u_2 \in \mathcal{S}$ if and only if u_1 and u_2 are friends on Facebook.

The Co-location Network: We construct an undirected graph based on user co-location, so that an edge exists between u_1 and u_2 if they were co-located. We call this graph the *co-location network* and denote it by $\mathcal{C}$.

The Co-located Friends Network: We will refer to the graph induced by those Facebook friends who were actually co-located as *the co-located friends network*, which will will denote by $\mathcal{S} \cap \mathcal{C}$.

Structural properties and descriptive statistics of the networks are shown in Table 1. In this work we will primarily focus on the edges of the graphs. Although their were 3636 observed colocations among the users, only 307 of these were co-locations of Facebook friends. This shows that although co-location among Locaccino users in Pittsburgh is quite common, co-location among Facebook friends is comparatively rare. Indeed, only roughly 30% of the dyads in the social network ever appear in the co-location network. Also of interest are global properties of the graph structures, such as the distribution of component sizes (see Figure 2). Observe that, ignoring isolated vertices, co-location of the participants occurs in one large connected component, whereas co-location of friends occurs in several smaller distinct components.

4 Model Descriptions

In this section we describe the variables we use to model the co-location of two users and individual user mobility.

4.1 Measuring the diversity of a location

To better understand the context of each observation, it would be helpful to have information about the type of location where the observation occurred. For example, observations of a user in a private residence should be interpreted differently from those in a crowded shopping center. We introduce a set of measures on locations that attempt to quantify the diversity of observations that occur at a given location. One primary motivation in defining these measures is to be able to distinguish when a co-location between two users happens by chance, say two strangers eating at adjacent tables in the same restaurant, and when the co-location is a social event, say one friend inviting the other to his house for dinner.

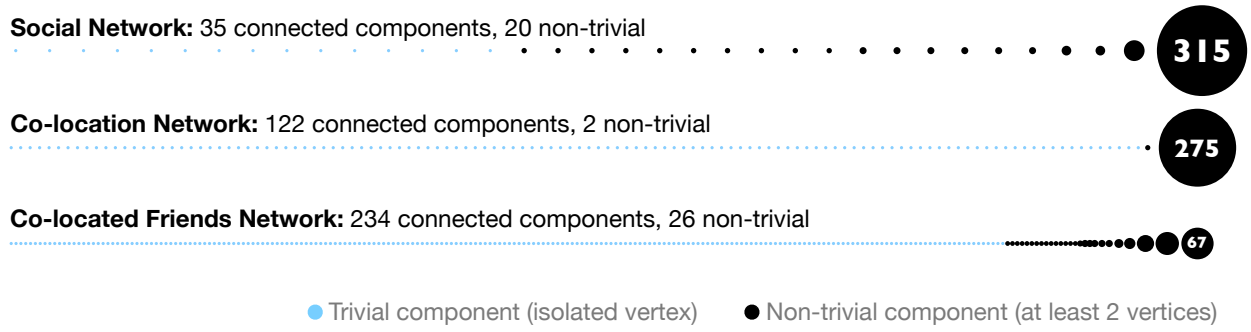

Figure 2: The distribution of connected component sizes in the social, co-location and co-located friends networks. Ignoring isolated vertices, co-location of the participants occurs in a single connected component, whereas co-location of friends occurs in several smaller components.

In this work we consider three diversity measure on location: *frequency*, *user count*, *and entropy*. The frequency of a location measures the raw count of user observations that occurred there, user count looks at the total number of unique users that visit the location, and entropy takes into account both the number of users observed at the location as well as the relative proportions of their observations. A location will have a high entropy if many users were observed at the location with equal proportion. Conversely it will have low entropy if the distribution of observations at a location is heavily concentrated on few users. See Figure 4 for a concrete illustration of the difference between the three diversity measures. We find entropy to be particularly appealing because locations of high entropy by definition are precisely where chance encounters are most likely to occur.

Now we define these notions formally. Let L be a location and let U be the set of all users. For a $u \in U$, let O_u be the set of location observations of u and let $O = \bigcup_{u \in U} O_u$. An observation $o \in O_u$ is a 4-tuple of the user ID, the location latitude and longitude coordinates, and a timestamp. Define $U_L = \{u \in U : u \text{ was observed at location } L\}$. Let $O_{u,L} = \{o \in O_u : o \in L\}$ and $O_L = \{o \in O : o \in L\}$ be restrictions of O_u and O to the location L[3]. The probability that a random drawn from O_L belongs u is $P_L(u) = \frac{|O_{u,L}|}{|O_L|}$, that is $P_L(u)$ is the total fraction of all observations at location L that are of user u.

Definition 1: For a location L, the *frequency* of the location is defined as $\texttt{Freq}(L) := |O_L|$, the *user count* of the location is defined as $\texttt{UserCount}(L) := |U_L|$, and the *location entropy* of the location is defined as $\texttt{Entropy}(L) := -\sum_{u \in U_L} P_L(u) \log P_L(u)$.

Our application of the `UserCount` and `Entropy` measures to study locations is motivated by their use in ecology in the study of biodiversity [13].

[3]The notation that $o \in O$ and $o \in L$ is imprecise. Elements of O are 4-tuples whereas elements of L are location coordinates. By $o \in L$ we mean the location component of o lies in location L.

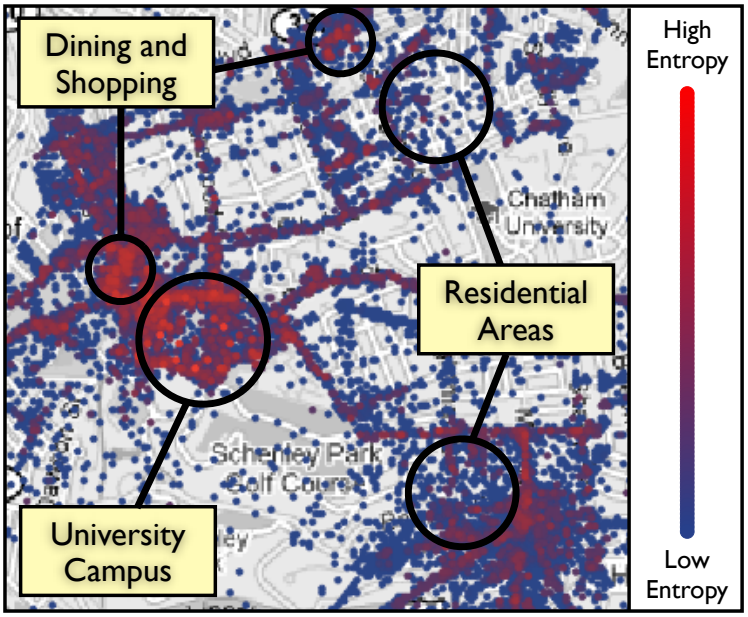

Figure 3: A map of the study region where observations are colored according to their level of entropy. Places such as the university campus, shopping and dining districts, have high entropy levels. Residential areas have low entropy.

4.2 Co-location features

For each co-location edge $\{u_1, u_2\}$ of $\mathcal{C}$, we extract 67 features from the data describing contextual properties of the history of co-locations between u_1 and u_2. These features, which are outlined in Table 2, are designed to distinguish more "meaningful" co-location histories from chance co-locations. Broadly, we divide the features into four categories, Intensity and Duration, Location Diversity, Specificity, and Structural Properties.

Intensity and Duration: The Intensity and Duration features measure qualities related to the size and spatial and temporal range of the set of co-locations. These features quantify how long and how actively users have embraced the system.

Location Diversity: The location diversity measures given in Definition 1 provide the basis for several features which aid in understanding the context of a set of co-locations. For a given co-location observation between u_1 and u_2, let l be the location where they were observed. We compute $\texttt{Freq}(l)$, $\texttt{UserCount}(l)$ and $\texttt{Entropy}(l)$ for every co-location of u_1 and u_2, then we take the average, median, variance, minimum and maximum of the resulting values to get the features listed in Table 2. Additionally, although they are not listed in Table 2, we also use two variations of each of these features where the statistics are taken only over evening and weekend co-locations.

Specificity: Inspired by the tf-idf ranking technique from information retrieval, we would like to

Category	Variables	Description	Co-location	User mobility
Intensity and Duration	NumObservations	The total number of observations of the user.		√
	NumColoc, NumColocEvening, NumColocWeekend	The number of co-location observations of the two users, in total, in the evening only, and on weekends only.	√	
	NumLocations, NumLocationsEvening, NumLocationsWeekend	The number of distinct grid boxes where the user or users were observed, in total, in the evening only, and on weekends only.	√	√
	NumHours, NumWeekdays, NumDates	The number of distinct hours of the day, days of the week, and calendar dates that the two users were observed together.	√	
	ObservationTimeSpan	The difference in seconds between the last and the first location or co-location observation.	√	√
	BoundingBoxArea	The area of the minimal axis aligned rectangle that contains the locations/co-location observations of the user/users.	√	√
Location Diversity	AvgEntropy, MedEntropy, VarEntropy, MinEntropy, MaxEntropy	The mean/median/variance/min/max of the location entropy at each location/co-location observation of the user/users.	√	√
	AvgFreq, MedFreq, VarFreq, MinFreq, MaxFreq	The mean/median/variance/min/max of the location frequency at each location/co-location observation of the user/users.	√	√
	AvgUserCount, MedUserCount, VarUserCount, MinUserCount, MaxUserCount	The mean/median/variance/min/max of the location user count at each location/co-location observation of the user/users.	√	√
Mobility Regularity	SchEntropyL, SchEntropyLH, SchEntropyLD, SchEntropyLHD	The schedule entropy of the user with respect to location, location and hour, location and day of the week, and location and hour and day of the week.		√
	SchSizeLH, SchSizeLD, SchSizeLHD	The schedule size of the user with respect to location and hour, location and day of the week, and location and hour and day of the week.		√
Specificity	AvgTFIDF, MinTFIDF, MaxTFIDF	The mean/minimum/maximum of the location TFIDF at each co-location of the two users.	√	
	PercentObservationsTogether	The total number of co-locations of the two users divided by the sum of each users total number of observations.	√	
Structural Properties	NumMutualNeighbors	The number of people who have been co-located with both users.	√	
	NeighborhoodOverlap	The number of people who have been co-located with both users divided by the number of people who have been co-located with either user.	√	
	LocationOverlap	The total number of distinct places visited by both users divided by the total number of places visited by either users.	√	

Table 2: Above are the names and descriptions of the independent variables used in our models. We divide the variables into five categories to better understand how the groups of variables relate to one another. Further we indicate whether the variables are features of co-location, or features of individual user mobility, or both.

measure how specific a location is to the pair of users who were co-located there. For example, a

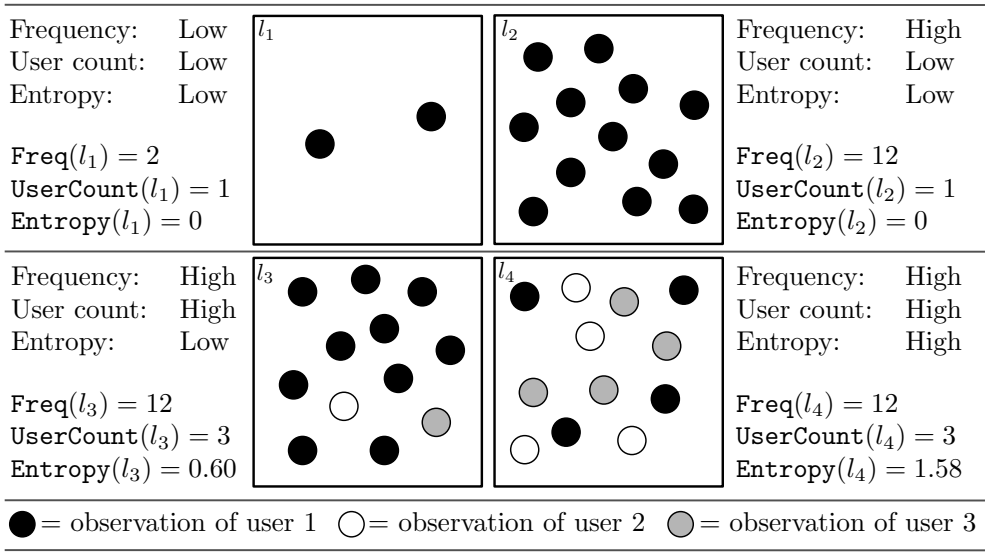

Figure 4: Four representative scenarios highlighting the differences between the location diversity measures. Circles of a given shade represent location observations of a particular user.

domestic residence is highly specific to the married couple that lives there, since a large fraction of the observations there are co-locations of that couple. For a given location l, we define the $\texttt{TFIDF}_{u_1,u_2}(l)$ to be the number of times u_1 and u_2 were observed co-located at l divided by $\texttt{Freq}(l)$ (i.e. the total number of observations at the location). The Specificity features listed in Table 2 are determined by first computing $\texttt{TFIDF}_{u_1,u_2}$ at each co-location observation of u_1 and u_2, and then taking the average, minimum, and maximum of the resulting data.

Structural Properties: We use three variables which measure the strength of the structural relationship between u_1 and u_2 in $\mathcal{C}$. Two of these are standard social network analysis techniques (NumMutualNeighbors and NeighborhoodOverlap). The third feature, LocationOverlap, is not strictly a structural property of $\mathcal{C}$, but is computed similarly to NeighborhoodOverlap, and can be viewed as a similarity measure between the sets of locations u_1 and u_2 visit.

4.3 Measuring the regularity of a user's routine

In studying how properties of user mobility relate to properties of the underlying social network, one attribute we would like to quantify is the regularity of a user's schedule. We accomplish this by first representing each location observation $o \in O_u$ as a vector of values of the location, day of the week, and hour of the day of the observation. To measure how a user's mobility pattern repeats in regular intervals, we can restrict this vector to a subset of the components and study the observed probability distribution in the resulting subspace.

For example, if we restrict the observations to just the location and day of the week components, then we could study how the user's schedule varies as a function of the day of the week. The

observed joint probability distribution on these two components provides some insight into the regularity of the user's weekly schedule. If the distribution is concentrated at relatively few values, then the user has a highly regular weekly schedule, and if the distribution is spread out among several values, then the user has a highly irregular schedule. We will again use entropy as a measure of the spread of this distribution.

We now formalize this intuition. Let $R \subset \{\text{L, D, H}\}$ denote the components of the restriction (standing for location, day of the week, and hour of the day respectively). For a given $o \in O_u$, we let $o(R)$ be the restriction of o to the components of S. Then $O_u(R) = \{o(R) : o \in O_u\}$ is the unique set of $|R|$-tuples generated by applying the restriction R to observation vectors in O_u (i.e. observations of users u). We define the observed probability distribution of the restricted observations as $r \in O_u(R)$ as $p(r) = \frac{|\{o \in O_u \,:\, o(R)=r\}|}{|O_u|}$.

Definition 2: The *schedule size* of u with respect to restriction R is defined as $\texttt{SchSize}(O_u, R) := |O_u(R)|$ and the *schedule entropy* of u with respect to R is defined as

$$\texttt{SchEntropy}(O_u, R) := -\sum_{r \in O_u(R)} P(r) \log P(r).$$

To clarify these definitions, suppose again that $R = \{L, D\}$, so R is a restriction of the observation to the location and day of the week. Then $\texttt{SchSize}(O_u, R)$ will be high if on each day of the week u visits many different locations and $\texttt{SchEntropy}(O_u, R)$ will be high if a u visits many locations on each weekday in relatively equal proportions.

4.4 User mobility features

For each vertex u of $\mathcal{C}$, we extract 64 features from the data describing properties of the mobility patterns of u. Again see Table 2 for a full description of the features. We partition the user mobility features into three categories, Intensity and Duration, Location Diversity, and Mobility Regularity.

Intensity and Duration: Similar to the corresponding category in the co-location model, these features measure the intensity of and range of the user's use of the system.

Location Diversity: Location Diversity features for the user mobility model are identical to the co-location model, except instead of measuring the diversity of co-location observations, we measure the diversity of the location observations of a single user.

Mobility Regularity: In this work we consider four restrictions: {L}, {L, H}, {L, D}, {L, H, D}. Computing the schedule size and schedule entropy on these four restrictions yields the seven Mobility Regularity features listed in Table 2 (the eight feature is $\texttt{SchSize}(O_u, \{L\})$, which is already represented by NumLocations). Similar to the location diversity variables, we also use evening and weekend variations for the Mobility Regularity features.

Classifier	Prec.	Recall
RandomForests (10 vars. per node)	0.62	0.22
RandomForests (18 vars. per node)	0.61	0.22
AdaBoost (dec. stumps, exp. loss)	0.68	0.24
AdaBoost (dec. stumps, lgstc. loss)	0.60	0.28
SVM (deg 2 polynomial kernel)	0.40	0.31
SVM (deg 3 polynomial kernel)	0.26	0.37

Table 3: The observed precision and recall of the 6 classifiers we tested. Predictions were conducted with a 50-fold cross validation procedure over all the observations in the dataset. The choice for number of variables at each split in the Random Forest models was picked via cross validation. Each Random Forest model had 1000 trees. The AdaBoost algorithm was run for 400 iterations.

5 Results

In this section we present our main results. We analyze the relative importance of the independent variables both in the context of predicting the number of social network ties a user has as well as the existence of a social network tie between two co-located users.

5.1 Inferring social network ties from co-location

First we consider the task of predicting ties from co-location data. For each edge of $\mathcal{C}$ we use a binary response variable FacebookFriends to indicate whether or not the corresponding edge is present in $\mathcal{S}$. In total, there were 307 co-location edges where the users were Facebook friends and 3330 co-location edges where the users were not Facebook friends. We model FacebookFriends as a function of the co-locations features.

We then trained 6 classifiers on the data (2 Random Forest classifiers, 2 AdaBoost Classifiers (with Decision Stumps), and 2 SVM Classifiers (with polynomial kernels). The performance of each classifier was measured against the true values of whether the users are Facebook friends using a 50-fold cross validation procedure. Table 3 shows the observed average precision and recall for each from this procedure.

The RandomForest models and the AdaBoost models outperform the two SVM models that we trained. In particular, the AdaBoost model with exponential loss seems to perform the best, having the best observed precision, and very near to the best observed recall. It correctly identifies 74/307 friendships and 3295/3330 non-friendships. Although the overall accuracy of the classifier is high (92%), this is somewhat misleading since the class distribution is heavily biased towards non-friendship, so we prefer to examine the precision and recall to get a clearer judge of performance.

To examine the relative predictive power of the 4 feature classes, we trained an AdaBoost classifier (exponential loss / decision stumps) using only the Intensity and Duration features. Such a model uses proximity-based features similar to those used by Eagle et al. for a similar friend prediction task [4]. To see how these feature perform against the location features we define in this work, we compare this model to one trained using the three remaining co-location feature classes

(Location Diversity, Specificity, and Structural Properties). The results are shown in figure 5. Here we plot the precision and recall curves at varying thresholds of the class probabilities output by AdaBoost. We again use 50-fold cross validation for all classifier estimates. For comparison, we also plot the precision/recall curves for the full AdaBoost classifier, and a baseline formed by thresholding NumColocations at varying threshold values.

One can observe that both the full model, and the model trained on the Location Diversity, Specificity, and Structural Properties features significantly outperform the model trained only on Intensity and Duration features. Furthermore at moderate to high recall levels, the Intensity and Duration model does not offer any improvement over simply thresholding on NumColocations. This is in contrast to the other models, which show consistent and significant gains over the baseline at all recall levels.

The low performance of the baseline and of the Intensity and Duration features is illustrative of the wide range of co-location patterns observed among the participants. This result shows that, although co-location alone is not a very strong predictor of online friendship, we can significantly improve the predictive performance by looking at additional contextual social properties of the locations the users visit.

5.2 Inferring the number of friends from user mobility data

Next we consider the relationship between the number of Facebook friends a user has in Locaccino and her mobility patterns. There are plausible hypotheses why one may expect variables in each of the three mobility feature categories to be correlated with the number of friends the user has in Locaccino. First, one would expect that user's who have used the system longer or more vigorously might have more friends in the system. Furthermore, since locations of high diversity are in some ways more "social," one could hypothesize that users who visit such locations often might have more friends in general. Finally, users with highly irregular schedules might find a system such as Locaccino more useful to help coordinate with their friend and family.

To examine this question, we first calculated the Pearson's correlation between the node degrees in $\mathcal{S}$ with each user mobility feature listed in Table 2. The results are plotted in Figure 6, where the three variable categories have been grouped and shaded in different colors. There are several interesting observations that should be noted when examining this figure. First, one should notice that the correlation of variables in the Intensity and Duration category have a far weaker correlation to the number of friends than the more nuanced variables in the Location Diversity and Mobility Regularity categories. Only 2 out of the 6 Intensity and Duration features correlate with the number of friends, and only very weakly (0.10 - 0.12).

The weak correlations of the Intensity and Duration variables is in contrast to the Location Diversity category, where the variable with highest correlation is MaxEntropyWeekend (cor=0.39 with 95% CI=(0.31, 0.47)). Indeed, MaxEntropyWeekend is the single variable most correlated to the number of friends. Furthermore, note that for each diversity measure, the highest correlated sample statistic on the set of locations is always the maximum: MaxEntropy (cor=0.29 with 95% CI=(0.20,0.37)), MaxUserCount (cor=0.30 with 95% CI=(0.21,0.38)), and MaxFreq (cor=0.29 with 95% CI=(0.20, 0.37)), with similar results for the evening and weekend variations of these variables. These variables each quantify in their own way the "most diverse" location

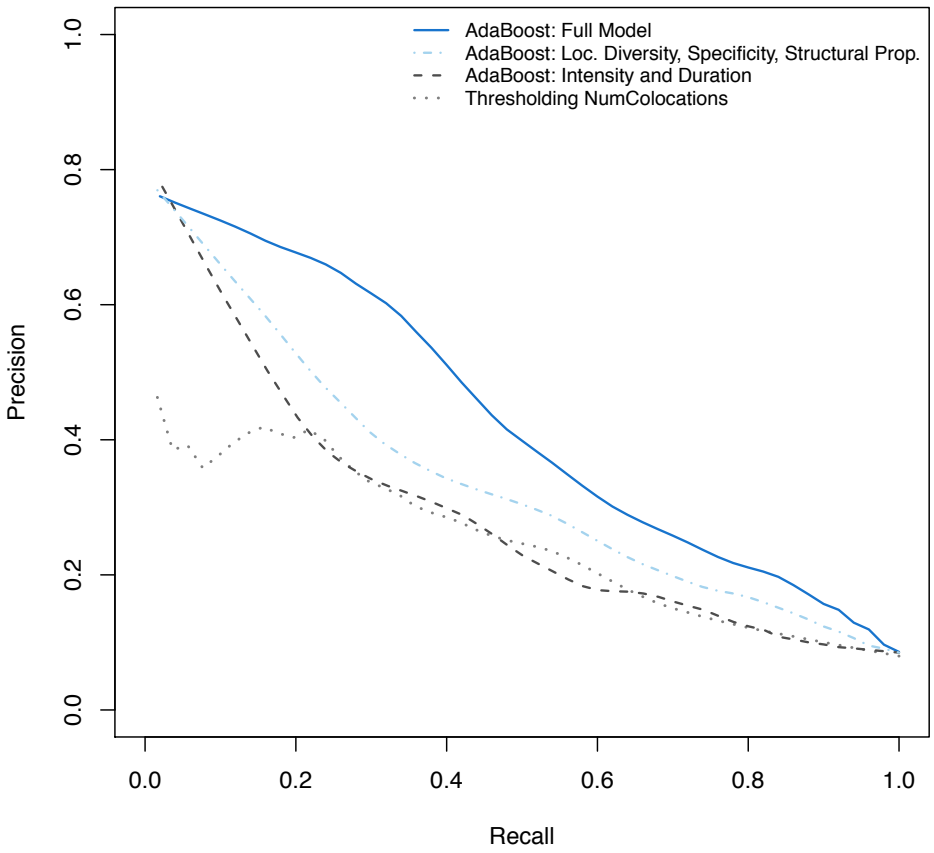

Figure 5: Observed precision and recall of the full AdaBoost classifier on all the features, compared with two sub-models: the first is only trained on the Intensity and Duration features, and the second is trained on the Location Diversity, Specificity, and Structural Properties features. Precision and recall estimates are made using 50-fold cross validation. These are shown against a baseline constructed by thresholding NumColocations at varying threshold values.

that a user visited, suggesting that users who visit highly diverse locations tend to have more social network ties than those who do not. In addition to the maximum, the average and variance sample statistics exhibit moderate positive correlations with the number of friends, whereas the minimum exhibits a weak negative correlation. The median statistic showed very weak and often insignificant correlations with the number of friends.

It is important to make the distinction that the location diversity measures for a user do not simply take census data of other nearby users at each specific location the user visits. Rather they are global properties of the locations themselves, taking into account all observations of all users at each given location (recall Figure 4). It is thus possible, even likely, for a user to be located at a highly diverse location, yet also not be co-located with any other system users. In this sense we believe these correlations are strong, if not surprising, results illustrating a unique relationship between the context of the locations a user visits and the number of online social network ties the

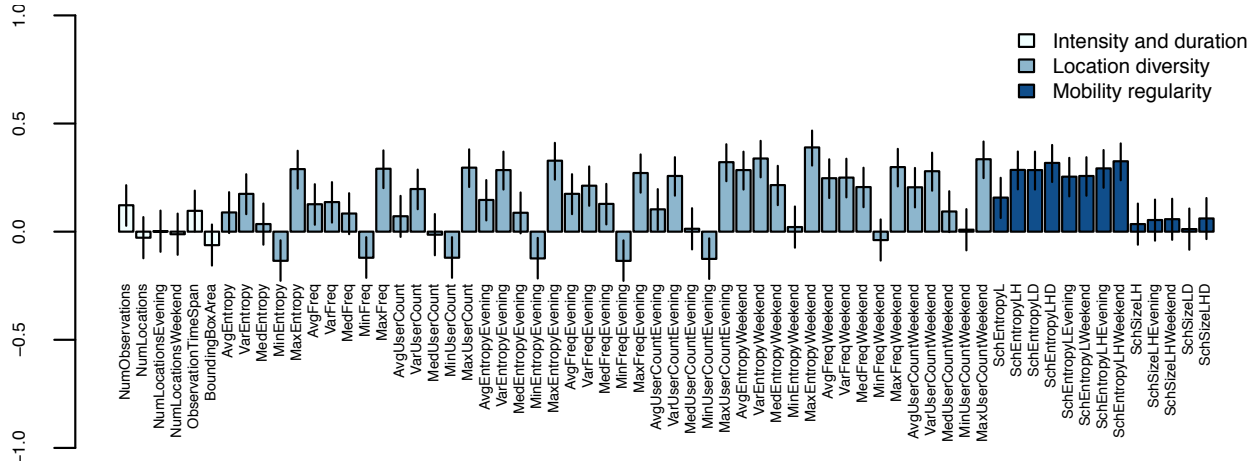

Figure 6: Pearson's correlation of user mobility variables with node degree in $\mathcal{S}$. Error bars indicate an approximate 95% confidence interval. The highest correlation values correspond to variables that measure the location diversity of observations.

user has.

Next observe the moderate positive correlation values for the schedule entropy variables in the Mobility Regularity category: SchEntropyL (cor=0.16 with 95% CI=(0.06,0.25)), SchEntropyLH (cor=0.28 with 95% CI=(0.20,0.37)), SchEntropyLD (cor=0.28 with 95% CI=(0.19,0.37)), and SchEntropyLHD (cor=0.32 with 95% CI=(0.23,0.40)), as well as their corresponding evening and weekend variations. As these variables quantify the regularity of a user's schedule, the results suggests that users who have irregular schedules tend to have more ties in the online social network $\mathcal{S}$.

It is notable that both the diversity variables and the regularity variables have a higher correlation with the number of friends than variables which measure the intensity and duration of system use. This suggests that the correlations observed in the diversity and regularity features are not simply byproducts of heavy system use.

To better understand the interrelations among the user mobility features with respect to the number of friends, we conducted a multiple regression analysis. First, to account for multicolinearity in the data, any pair of independent variables having correlation higher than 0.80 in absolute value, the variable with lowest absolute correlation to the number of friends was discarded. We then performed an stepwise search with AIC penalty working backwards from the full model to select a sub-model of the full linear model. The fitted model of the remaining 10 variables (2 intensity, 7 diversity, 1 regularity) given by this procedure are shown in table 4.

The resulting model (adj R^2=0.21, p-val < 0.001) yields very strong evidence that the diversity and regularity variables outperform the intensity variables. Examining the standardized β coefficients from the regression can provide some insight into which variables are the strongest predictors. We can see that the two variables with highest absolute β are AvgFreqEvening ($\beta = 0.223$) and SchEntropyLHWeekend ($\beta = 0.237$). Indeed, the Intensity and Duration variable in total

Variable	b-estimate	β-estimate	p-value
SchEntropyLHWeekend	7.67e-01	0.237	0.001
AvgFreqEvening	1.25e-04	0.228	< 0.001
MinFreqEvening	-1.83e-04	-0.195	0.002
MaxEntropyWeekend	6.99e-01	0.188	0.007
VarEntropyEvening	8.46e-01	0.148	0.008
MinEntropy	9.08e-01	0.119	0.055
BoundingBoxArea	-2.97e+02	-0.102	0.037
VarFreq	-3.01e-09	-0.092	0.098
MinFreqWeekend	-4.26e-05	-0.042	0.398
NumObservations	-3.32e-05	-0.037	0.500

Table 4: Raw b-coefficient, standardized β-coefficient estimates and p-values from a multiple regression with listed variables taken independents and the number of friend taken as dependent. Variables are sorted according to the absolute value of the β estimates.

comprise 10% of the total absolute weight of the β coefficients, whereas the Location Diversity variables comprise 73%, and the Mobility Regularity terms comprise 17%.

This result provides evidence that an examination of the context of the locations a user visits and analysis of the regularity of the user's routine can provide valuable insight into social behaviors of the user, in this case the number of friends the user has in an online social network. We have shown that the types of places a user visits, and the regularity of a user's routine are stronger predictors for the number of Locaccino friends they have than how long or how intensely they use the system.

6 Discussion

In this work we have explored several interesting connections between an online social network, and an offline co-location network on the same user set. These networks have very different structures. The co-location network has roughly 3 times the number of edges as the social network, yet the social network is better connected. The co-location network has many small disconnected components, but it has a single large and highly connected subcomponent. Despite these differences, we have shown that the co-location graph contains important information that can be used to reconstruct a portion of the social network.

We have shown that properties of the locations a user visits can provide valuable context to the user observations. In particular we have shown that the entropy of a location is a valuable tool for analyzing social mobility data. By definition, locations of high entropy locations are precisely the places where chance encounters are most probable, thus co-locations at high entropy locations are thus much more likely to be random occurrences than co-locations at low entropy locations. Thus if two users are only observed together at a locations of high entropy such as a shopping mall or a university center, they are less likely to actually have a tie in the online social network than if they

are observed in a place of low entropy.

We have also shown that the entropy of the locations a user visits can provide insight into the number of ties that the individual has in the social network. Users who visit locations of higher entropy tend to have more ties in the social network than users who visit less diverse locations. One possible explanation for this result is that locations of high entropy tend to be more social in nature than locations of lower entropy, and so users who visit these locations tend to be more social. However future studies are needed to further explore this relationship.

In addition to location diversity measures, we have explored several novel features that have proven useful in analyzing social mobility data. We looked at features that measure the intensity, location diversity, specificity, and structural properties of a set of co-locations, and we used these to construct a classifier that predicts social network ties between the users. These features far outperform predictions based on simple co-location observation counts between the two users.

In our analysis we have observed a wide range of co-location patterns between both Facebook friends and non-friends. We view this as testament to the complexities of human social relations (both online and offline). Indeed, the data show many instances where users are not friends in the online social network, yet exhibit very convincing co-location patterns for friendship. Similarly, there are numerous instances of friendships in the online social network with little to no evidence for friendship in the co-location data.

This disparity highlights two strong use cases for our work. Online social networks could use our classifier in their friend recommendation systems to find users with strong co-location patterns who are not yet friends in the social network. Such a system could strengthen current link-based friend recommendation systems by taking into account user behavior in the offline world to bolster online social relationships. Additionally, although future research is needed to verify the hypothesis, it is plausible that the predictions (and mis-predictions) of our classifier could provide insight into the strength of ties between users [8, 6]. If this were true our work could aid location aware social networks in developing systems to aid users in segregating and categorizing their online connections, which among other things could be useful in building privacy rules and organizing the social graph.

One area where we feel our work has the greatest potential is as a window into the relationship between online and offline social behavior. We show that location-based features (such as the entropy of a location) have significant correlations with real social behavior features (such as the number of friends in a social network). Understanding the interplay between users' location patterns and social patterns is an important area for future research.

It is also important to highlight some of the limitations of our work. Location data is extremely sensitive [14], and it is clear that the type of analysis we perform requires strong privacy controls and procedures that would protect users. Also, our pool of study participants is highly homogenous, containing mostly students. Future studies would strengthened by seeking a more diverse pool of participants as other populations could exhibit different online and offline behavior.

7 Conclusions

In this work we explore connection between an online social network and the location traces of its users. We evaluate a set of features of the location observations for their potential in analyzing the social behavior of the users. Social network designers may find our methodology useful for designing social applications, such as location-aware information sharing platforms, privacy control mechanisms, and friend suggestion systems.

This work opens up many future paths of research. Can different types of social relationships be inferred from location data? Can tie strength be estimated from locations? Does offline interaction spur online communication? This also raises important privacy questions about how much information location-based services leak about their users. We believe that this work provides a necessary step towards addressing such questions.

8 Acknoledgements

This work is supported by NSF Cyber Trust grant CNS-0627513, by CyLab at Carnegie Mellon under grant DAAD19-02-1-0389 from the Army Research Office, and by the Fundação para a Ciência e a Tecnologia through the Carnegie Mellon Portugal Program. Additional support has been provided by Microsoft through the Carnegie Mellon Center for Computational Thinking, and by grants from France Telecom, Nokia and Google. The authors would also like to thank David Eggerschwiler and Jay Springfield for developing the locator software, and the members of the Mobile Commerce Lab for their feedback.

References

[1] CULP, M., JOHNSON, K., AND MICHAILIDES, G. ada: An r package for stochastic boosting. *Journal of Statistical Software 17*, 2 (9 2006), 1–27.

[2] DERESIEWICZ, W. Faux friendship. *The Chronicle of Higher Education* (2009).

[3] EAGLE, N., AND PENTLAND, A. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology 63*, 7 (May 2009), 1057–1066.

[4] EAGLE, N., PENTLAND, A. S., AND LAZER, D. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences 106*, 36 (September 2009), 15274–15278.

[5] ELLISON, N. B., STEINFIELD, C., AND LAMPE, C. The benefits of facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication 12*, 4 (2007).

[6] GILBERT, E., AND KARAHALIOS, K. Predicting tie strength with social media. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems* (New York, NY, USA, 2009), ACM, pp. 211–220.

[7] GONZÁLEZ, M. C., HIDALGO, C. A., AND BARABASI, A.-L. Understanding individual human mobility patterns. *Nature 453*, 7196 (June 2008), 779–782.

[8] GRANOVETTER, M. S. The strength of weak ties. *The American Journal of Sociology 78*, 6 (1973), 1360–1380.

[9] HAMPTON, K., SESSIONS, L., HER, E. J., AND RAINIE, L. Social isolation and new technology. Tech. rep., Pew Internet and American Life report, November 2009.

[10] KRAUT, R., PATTERSON, M., LUNDMARK, V., KIESLER, S., MUKOPADHYAY, T., AND SCHERLIS, W. Internet paradox: A social technology that reduces social involvement and psychological well-being. *American Psychologist 53* (1998), 1017–1031.

[11] LI, Q., ZHENG, Y., XIE, X., CHEN, Y., LIU, W., AND MA, W.-Y. Mining user similarity based on location history. In *GIS '08: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems* (New York, NY, USA, 2008), ACM, pp. 1–10.

[12] MIKLAS, A. G., GOLLU, K. K., CHAN, K. K. W., SAROIU, S., GUMMADI, K. P., AND DE LARA, E. Exploiting social interactions in mobile systems. In *UbiComp'07: Proceedings of the 9th international conference on Ubiquitous computing* (Berlin, Heidelberg, 2007), Springer-Verlag, pp. 409–428.

[13] RICOTTA, C., AND SZEIDL, L. Towards a unifying approach to diversity measures: Bridging the gap between the shannon entropy and rao's quadratic index. *Theoretical Population Biology 70*, 3 (2006), 237–243.

[14] SADEH, N., HONG, J., CRANOR, L., FETTE, I., KELLEY, P., PRABAKER, M., AND RAO, J. Understanding and capturing peoples privacy policies in a mobile social networking application. *Journal of Personal and Ubiquitous Computing 13*, 6 (August 2009).

[15] WELLMAN, B., HOGAN, B., BERG, K., BOASE, J., CARRASCO, J.-A., CÔTÉ, R., KAYAHARA, J., KENNEDY, T. L. M., AND TRAN, P. *Networked Neighbourhoods*. Springer, 2006, ch. Connected Lives: The Project.

[16] WYATT, D., BILMES, J., CHOUDHURY, T., AND KITTS, J. A. Towards the automated social analysis of situated speech data. In *UbiComp '08: Proceedings of the 10th international conference on Ubiquitous computing* (New York, NY, USA, 2008), ACM, pp. 168–171.

[17] ZHENG, Y., LI, Q., CHEN, Y., XIE, X., AND MA, W.-Y. Understanding mobility based on gps data. In *UbiComp '08: Proceedings of the 10th international conference on Ubiquitous computing* (New York, NY, USA, 2008), ACM, pp. 312–321.