Modeling the Co-evolution of Behaviors and Social Relationships Using Mobile Phone Data

Wen Dong[1], Bruno Lepri[1,2] and Alex (Sandy) Pentland[1]
[1]MIT Media Laboratory
Cambridge, MA 02142, USA
[2]FBK
Povo-Trento, Italy
Email: {wdong, pentland}@media.mit.edu, lepri@fbk.eu

ABSTRACT
The co-evolution of social relationships and individual behavior in time and space has important implications, but is poorly understood because of the difficulty closely tracking the everyday life of a complete community. We offer evidence that relationships and behavior co-evolve in a student dormitory, based on monthly surveys and location tracking through resident cellular phones over a period of nine months. We demonstrate that a Markov jump process could capture the co-evolution in terms of the rates at which residents visit places and friends.

Categories and Subject Descriptors
H.1.2 [Models and Principles]: User/Machine Systems – Human information processing.

General Terms
Algorithms, measurement, experimentation, human factors.

Keywords
Social computing, human dynamics, living lab, stochastic process, multi-agent model.

1. INTRODUCTION
People act, meet, and interact in the context of time and space. As such, time and space provide important hints about the co-evolution of individual behavior and social interaction – people change where they go, and when they go there, by chance and through social interactions, and make connections and influence one another through their shared time and space. This paper discusses the spatial-temporal patterns of the residents in an MIT student dormitory over nine months, tracked every six minutes through the cellular phones of the students, as well as through our dynamical-process approach. With these data we capture the patterns of the students and make inferences regarding co-evolution. The patterns reported by this paper are generalizable, and our method could serve as the inference engine of cell-phone applications to integrate sensor data.

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.
MUM'11, Dec 7–9, 2011, Beijing, China.
Copyright © 2011 ACM 978-1-4503-1096-3/11/12…$10.00.

Time and space have important practical implications in shaping the behavior and social networks of individuals, and in modeling the co-evolution of individual behavior and social interaction. For example, let us assume that health is related to where people go and how they interact with one another. We can construct cell-phone applications to track this behavior and social interaction, and can subsequently improve health by showing Joe how his behavior and social interaction could affect his fitness, and by offering Joe rebate on health insurance and on the subscription fees of fitness centers if his cell-phone indicates that he commits to regular physical exercises in fitness centers with his neighbor Jane. In this way, Joe gets improved fitness and would like to pay for his improvement, the fitness centers get more customers revenue around an athletic culture, and the insurance companies get lower risk on their customers.

Due to the lack of time and space data, the co-evolution of individual behavior and social interaction is not well understood. For example, Christakis et. al. [1][2][6] claimed that obesity, happiness, and smoking behavior are contagious, based on 32 years of medical records involving 12067 people considered together with their relatives and emergency contacts. On the other hand, there have recently been concerns [14] about whether these medical records were sufficient to properly establish this contagiousness, and whether the claimed contagion could be otherwise explained by people with similar behaviors preferring to spend time together. Closely tracking the locations and proximities of people in communities over a span of months enables us to better determine the relationships between changing behavior and changing social network – whether strangers of the same "feather" are more likely to share time and space, and whether friends have increasingly-similar schedules. The causal relationship in dispute has practical implications.

If individual behavior is contagious, then we can change this behavior either by changing the behavior of several influential elements in the social network, or by changing the social network itself. If individual behavior is not contagious and people with similar behavior simply go together like "birds of a feather," then we need not bother with social networks in shaping individual behavior.

Previous attempts in untangling the problem of the co-evolution dynamics without closely tracking the everyday life of a complete community has been inconclusive, and tracking behavior using mobile phones offers new hope. Kossinets and Watts [12] analyzed a dynamic social network comprised of 43553 students, faculty, and staff at a university, in which the social interactions were characterized by who sends emails to whom and when, and

individuals were characterized by their gender, age, department, number of years in the community, and classes taught or taken. The researchers found that shared time and space (in this case, classes) and shared acquaintances increase by more than ten times the chance of forming a social tie. In comparison, attributes such as status, gender and age do not affect the chance of forming a tie [8]. Tracking data could reveal how acquaintance increases the chance of forming a tie, which other shared times and spaces besides classes are important, and whether shared time and space is the cause of a tie, or the effect.

We conducted the Social Evolution experiment to closely track the everyday life of a whole community with mobile phones, so that the social science community can validate their models against the spatio-temporal patterns and behavior-network co-evolution as contained in this data. The Social Evolution experiment covered the locations, proximities, and phone calls of more than 80% of residents who lived in the dormitory used in the Social Evolution experiment, as captured by their cell phones from October 2008 to May 2009. This dormitory has a population of approximately 30 freshmen, 20 sophomores, 10 juniors, 10 seniors and 10 graduate student tutors. -We conducted surveys monthly on different social relationships, health-related issues and statuses, and politics-related issues and statuses. This experiment also captured the locations and proximity of residents through a cell-phone application that scans nearby Wi-Fi access points and Bluetooth devices every six minutes – referenced to the latitudes and longitudes of the Wi-Fi access points – and then compared them to student demographics data to make sense of the data set. The data are protected by MIT COUHES and related laws.

This paper contributes to the discourse of computational social science in several ways. We are the first to report via information captured by cell phones where students go and how they interact, correlated through survey results with how their locations and interactions are related to important issues such as relationships and health. The dynamics reported in this paper are likely to be generalizable because they are compatible with previous findings. We introduced the dynamical process model to describe how people with different attributes (for example, years in school, or amount of physical exercise per week) go to different places, and how people with relationships co-occur in space and time. Such a model is useful in several ways. As a result, we can now sample "typical" time series of behavior and interactions from the model when sensor data is missing, or when the real sensor data cannot be published due to privacy concerns. We can also design the most efficient way to collect information and to protect privacy on a limited budget. Overall, from these data we can simulate different ways to shape behavior and relationships.

The rest of this paper is organized as follows: Section 2 revises some previous and related works. Section 3 describes the Social Evolution data and in particular Section 3.1 describes how residents with similar attributes and behaviors form relationships, and conversely how their relationships shape their attributes and behaviors in the survey. Section 3.2 describes how the co-evolution of individual behaviors and social interactions from the survey results is related to the sensor data. Section 4 describes our dynamical process model. Section 5 puts the spatiotemporal dynamics into applications in order to infer individual behaviors and social relationships, and discusses data collection and privacy. Finally, Section 6 draws some conclusions and proposes some perspectives for future works.

2. PREVIOUS AND RELATED WORKS

Previously, researchers have tracked individual behavior and social interactions in 100-person communities using mobile phones [5], motion sensors [24], and sociometric badges [1]. Researchers have also investigated the individual behavior and social interactions in geographical regions of one thousand to one hundred million people by analyzing cell-phone service provider call records [8] and vehicle-tracking records [11]. These studies reported about periodicity in human behavior, or social network structure, or the centrality-productivity relationship. However, the usability of such sensor data is restricted by the limits on collecting data for additional investigations beyond the original purposes of the data.

Of the previous data collected, the Reality Mining data set most resembles the Social Evolution data set in terms of tracking with cell phones the location, proximity, and other behaviors of several connected research groups [5]. The Reality Mining data have lower spatial resolution than the Social Evolution data, however, because the former used cell towers to represent locations while the latter used Wi-Fi access points, which are more closely clustered. The Reality Mining data also contain fewer and less-frequent survey records than the Social Evolution data. The Reality Mining data demonstrated the potential for modeling group dynamics by closely tracking locations, proximity, and other phone-recordable activities, compared to traditional methods such as conducting surveys. That experiment showed that a subject's satisfaction with his workgroup is dependent on whether he spent time with other persons from the workgroup during evenings and weekends. If this pattern is generalizable, a cell-phone application could improve the satisfaction of workgroup members either by placing people in the most appropriate workgroup or by facilitating off-hour socialization. The experiment also showed how new members gradually adopted workgroup-specific social norms, such as work hours and interactions with other groups. A cell-phone application could consequently be useful for preserving workgroup-specific social norms and for assisting with the adaptation of new workgroup members. In these scenarios, a survey-based approach is clumsy compared to a sensor-based approach.

Of the previous models on community dynamics, the exponential random graph (ERG) models most resemble the Markov jump process model proposed in this paper. ERG models have received significant attention in the past few years in computational social science. In a nutshell, ERG models describe parsimoniously how local selection forces (attributes of individual nodes and dyads) shape the global structure (e.g., centrality distribution and degree distribution) of a network [7][10][13][23]. We use Markov jump process to model community behavior because at the temporal resolution of the sensor data (several minutes), individual behavior and social interaction do not have enough time to attain maximum entropy distribution, as is required by the temporal ERG model. A Markov jump process has wide applications in modeling processes of discrete events in finance, systems biology, chemistry and physics.

3. OUR DATA

The dataset used in our experimental analyses was collected as part of a longitudinal study with seventy residents of an undergraduate residence hall (referred to as an undergraduate dormitory in North America), which serves as the primary residential, cooking, social, and sleeping quarters for residents. This residence hall is the smallest undergraduate dormitory at the university. Participants in the study represent 80% of the total population of this hall, and

most of the remaining 20% are spatially isolated. This dormitory is known within the university for its pro-technology orientation, and residence in the dorm is determined through self-selection by both applicants to the dorm and the existing residents. The students were distributed roughly equally across all four academic years (freshmen, sophomores, juniors, seniors). 54% of the students were male, and predominantly engineering, mathematics, and science majors. The study participants also included four graduate resident tutors who supervised each floor. Participants used data-collecting Windows Mobile devices as their primary phones, with their existing voice plans. Students had online data access with these phones due to pervasive Wi-Fi on the university campus and in the metropolitan area. As compensation for their participation for the entire academic year, participants were allowed to keep the smartphones at the end of the experiment. Additional information is available in Madan et al. [16][17][15].

The dataset used in this analysis was collected from (i) self-reported surveys conducted monthly and designed by experts in political sciences and medicine, and (ii) cell-phone sensors that record proximity and location every six minutes and document communication. Data from surveys include the relationships of close friends; with whom residents are socializing, discussing politics, sharing Facebook photos, and sharing Twitter and blog posts; and attributes about individuals, such as participation in on-campus organizations, health conditions and involvement in physical exercises, political opinions and involvement, and demographic information. Data from sensors also include dyadic relationships such as who called whom, who sent short messages to whom, physical proximity; and attributes about individuals, such as Wi-Fi hotspot scanning, which tracks where the individuals were.

3.1 Survey Data

The shared time, space, and relationships are the foundations from which the residents in the dormitory built additional relationships, as suggested by our monthly relationship surveys. The shared living sector in the dorm was the most important factor, especially for new residents. Shared courses and shared on-campus extra-curricular activities were also important factors.

The shared living sector is the most important factor for a student to build relationships. We collected the room numbers of 71 of the 84 residents, sharing eight living sectors separated by the four floors and a firewall. A given student was five times more likely to report another student in his sector as a friend than another student living in a different sector. The 18 surveyed residents living in double rooms all reported their roommates as friends. At the beginning of a school year, the average student socialized with half of the students living in his sector, and with only 2-3 of students living in different sectors. At the end of a school year, he socialized with about one-third of students in the same sector, and with only 1-2 students in other sectors.

The students' academic and extra-curricular activities were also important contexts for building relationships. The average student was five times more likely to report as a friend another student in the same year at school. The average numbers of friends reported by freshmen, sophomores, juniors, seniors, and graduate tutors were respectively 2.9, 7.1, 8.1, 5.4 and 4.7 in the first month of the school year. Students were less likely to report each other as friends if their year in school differed by more than one. Over time, every freshman made five friends on average, while every graduate tutor made up to nine friends. Residents who had already stayed in the dormitory for more than one year (sophomores, juniors, and seniors) changed less than 10% of their friendship relations.

We speculate that friendship influenced the assignment of living sectors, because freshmen were distributed among the eight living sectors more randomly than other residents. (The entropy of freshman distribution among the eight living sectors is 1.8, while the entropies of the 2nd, 3rd, 4th-year students are all around 1.6.) Let us adopt the mild condition that the statistics of the freshmen were the same as the statistics of the sophomores one year before the Social Evolution experiment (when the sophomores were freshmen), and the statistics of the juniors two years before the experiment, and so on. Let us adopt the further another mild condition that the probability distributions under our inspection have maximum entropy among all distributions compatible with our observations – that is, no one can deliberately arrange the data in a specific way to defeat our reasoning. It then follows that freshmen change their sectors to live in concentrations at the end of a school year. Since those who are the same year in school are more likely to be friends, we conclude that friends are more likely to change to live in a concentrated environment.

Figure 1 shows how the friendship network evolves from September 2008 to March 2009. Different living sectors are represented with different colors. An opening in the firewall connects dormitory sectors "f282.3" and "f290.3." Numbers 1 to 5 represents freshman, sophomore, junior, senior and graduate resident tutor respectively. The relationships in September 2008 were based mostly on living sectors, because the freshmen met mostly those in the same sector, and the other residents had already adjusted their living sectors based on existing relationships. New relationships between September 2008 and March 2009 were mostly connected with the freshmen.

As such, the data suggest that friendship in a student dorm are based on common experiences of life and other friendships, which in turn shape experiences of life, where "experience" was

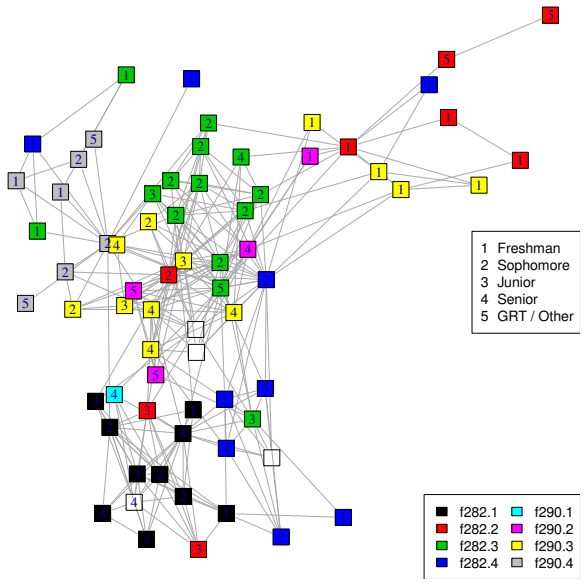

Figure 1: Subjects in the dormitory formed clusters of relationships by their dormitory sectors (the primary factor) and their years in school (the second most important factor).

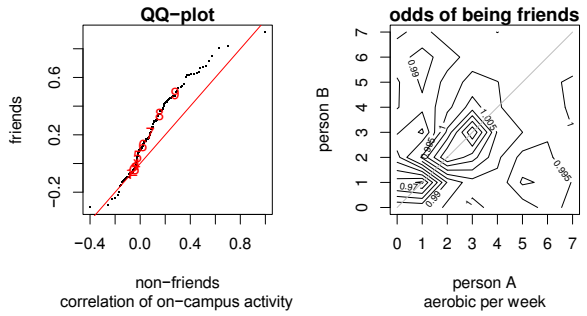

Figure 2: A pair of friends has higher correlation in their on-campus activity participation than a pair of non-friends.

identified from surveys as common living sectors and curricula. This interaction between friendship and behavior coincides with findings elsewhere.

Friends also have a higher correlation in their on-campus activities, as indicated by the monthly surveys. From September 2008 to May 2009, 15 friend pairs each shared all on-campus activities that we surveyed, and 30% of friend pairs shared over 50% of their on-campus activities. In comparison, non-friends shared less than 10% of on-campus activities. In particular, pairs who participated aerobic exercise about three times weekly were more likely to be friends. While the surveys show significant correlation between activity participation and friendship (hypothesis that the correlation of friends' activity participation has the same probability distribution as the correlation of non-friends' was rejected with $p < 10^6$ in a Kolmogorov-Smirnov test), activities and interactions collected by sensors are nonetheless necessary for estimating friendship or activities, as the surveys do not offer statistically-significant correlations between friendship and factors that do not require shared space and time, such as shared websites and shared music.

The left panel of Figure 2 compares activity participation between friend pairs and non-friend pairs with a quantile-quantile plot (QQ plot). A QQ plot compares the probability distributions of two samples. When the two samples (participation correlations among friends and participation correlations among non-friends) have equal sizes, the QQ plot draws in a coordinate system the lowest value in one sample (non-friends) against the lowest value in the other sample, draws the second-lowest value in one sample against the second-lowest value in the other sample, and so on. When the two samples have different sizes, the QQ plot interpolates the values in the two samples. The digits 1-9 in this panel mark the 0.1-0.9 percentiles in the two distributions. For example, 20% of friend pairs and 5% of non-friend pairs (marked by red digit 8) have activity correlations greater than 0.4. If we identify friends as those whose activity correlations are greater than 0.4, we can successfully identify the 20% of friend pairs with the highest activity correlations, but also misidentify the 5% of non-friend pairs with the highest activity correlations as friend pairs.

The right panel of Figure 2 shows the odds that two individuals will be friends given that they perform certain numbers of aerobic exercises per week. When both individuals participate in around three aerobic activities per week, they have higher odds of being friends. This suggests that friendship determines behavior, because otherwise it is hard to explain why people who do 2.5 aerobic

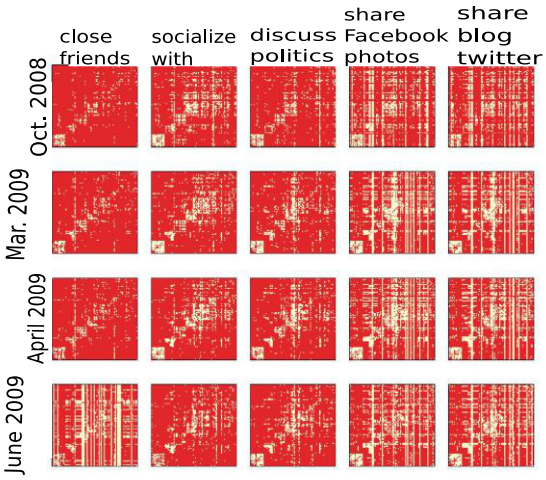

Figure 3: Students in the dormitory reported clusters of relationships in the surveys.

activities per week like those who do 2.5 aerobic activities per week, but people who do 3.5 aerobic activities per week like those who do 3.5 aerobic activities per week. When both individuals have one and fewer aerobic activities per week they are more likely to be non-friends. This seems to suggest that physical exercise is an additional factor other than shared living sector and shared courses that also shape friendship relations.

Figure 3 shows each socialization network, political discussant network, the Facebook network, the blog/Twitter network (columns 1-5) during December 2008, March 2009, April 2009 and May 2009 (rows 1-4). Each of the 20 images shows whether a given person A (x-axis) reported the stated relationship with a given person B (y-axis) in a specific month. We have reordered the participants so that friends go together. The friendship networks, the socialization networks, and the political discussant networks all take the block diagonal form, and the blocks represent different living sectors. Hence, relationships and living sectors have the most important interactions. The Facebook networks and blog/Twitter networks are the least structured. Some individuals reported relationships with all other residents in the student hall, especially towards the end of the academic year. This may indicate that by the end of the year all 84 residents in the small student dorm know one another, and distinguishing between the different relationships in the survey is a difficult task as a result.

3.2 Sensor Data

While the monthly surveys offer important insights into the co-evolution of relationships and behavior, the cell-phones applications record relationships and behavior in much deeper detail, and provide almost limitless data-mining potential. The behavior and interactions recorded by the cell phones show clear daily, weekly, and yearly patterns, and increased behavioral complexity the longer a resident stays in school. There are positive correlations between reports from surveys and data from the cell phones; as such, the phones provide a way to fill in details between the surveys, and to potentially understand the behavior-interaction co-evolution.

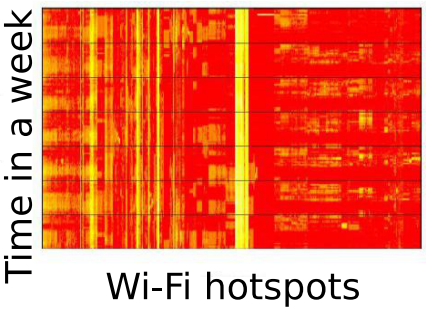

Figure 4: The student dormitory community cycled among dormitory (left), athletic center (middle) and classroom/office (right) from Sunday (bottom stripe) to Saturday (upper stripe), as indicated by Wi-Fi access-point usage.

We make sense of the Wi-Fi data by collecting information about the access points – their latitudes and longitudes, and what activities people normally do around them – and use web-mapping technologies to visualize the co-evolution of behavior and interactions in time and space. This information from Wi-Fi access points and web mapping technologies was greatly helpful for us in making inferences and validating our models.

Figure 4 is a heat map showing how individuals visited places daily. The x-axis is indexed by Wi-Fi access points, and the y-axis is indexed by time during the week from Monday morning to Saturday at midnight. An entry shows how often the residents accessed a Wi-Fi access point in a specific hour in the week. The Wi-Fi access points on the left side were in the dormitory building, and so had many accesses from midnight to morning. The Wi-Fi

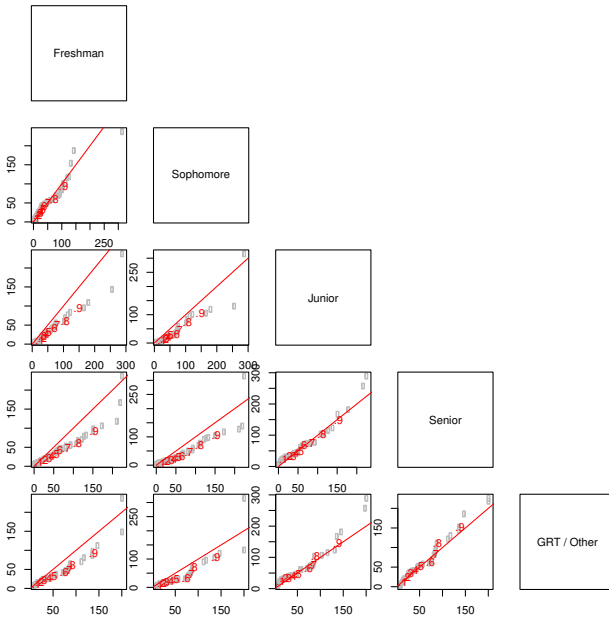

Figure 5: Students frequented more locations as they stayed in school longer.

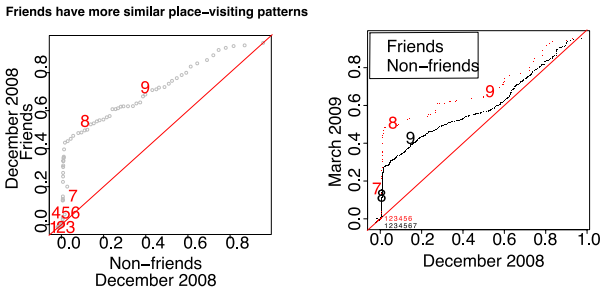

Figure 6: Friends have similar location-visiting patterns (left), and increase in location-visiting similarity over time (right).

access points on the right had high usage during work hours, and correspond to the classrooms and offices. The Wi-Fi access points in the middle show high usage from evening to midnight, and correspond to fitness centers and the student activity center.

More years students had been in school, more Wi-Fi access points they frequented. Since MIT is covered with Wi-Fi access points, and since each location on average is within the range of 10 such points, we infer that over the years in school students add new places to their repositories of familiar and frequented locations, and so increased the complexity of their on-campus activities. Figure 5 uses QQ plots to compare the numbers of frequently-encountered Wi-Fi points for the freshmen, sophomores, juniors, seniors, and graduate student tutors. The freshmen averaged 50 frequently-visited Wi-Fi access points, the equivalent of 5 distinct locations. The sophomores begin to demonstrate a trend of adding new places to frequent, and the juniors and seniors demonstrate an average of 80 frequently-visited Wi-Fi access points, the equivalent of 8 locations. To compute the number frequently-visited places, we first count how many times a student visited each of the Wi-Fi access points that he visited, then compute the entropy (the average of the logarithm of the visiting counts), and use the exponential of entropy as an indicator of the number of frequently visited places.

Friends "infect" one another in terms of how frequently they visit places. Hence, friendship is an important contributor to the increased number of frequently-visited places. This infection manifests itself as friends visiting places with increasingly-similar frequencies over time. On the other hand, two people are also more likely to become friends when they share spaces for significant amounts of time (Figure 6). We argue that shared space shapes friendship, since the higher correlation of space usage among freshmen (who didn't know one another before the experiment) cannot be otherwise explained.

Figure 7 shows the odds that two persons became friends, given that they met at a specific longitude and latitude. Only at a few locations – such as the dormitory, office buildings, classrooms, cafeteria, and athletic center – were two students more likely to be friends. In other locations, any person encountered was more likely to be a non-friend. This can be explained by the fact that friends often stay together longer at meaningful places, while the non-friend pairs often just passed by one another.

Figure 7: Friends often met at a few meaningful places. In contrast, non-friends passed each other randomly.

4. OUR MODEL

We use the Markov jump process to model sensor events regarding location and proximity related to the reported behaviors and interactions in the surveys. The events of this Markov jump process model are the time-stamped records from Wi-Fi hotspot access, Bluetooth scans, SMS's and phone calls. The rates of different events are parameterized by the surveyed individual behavior and interpersonal relationships. Using the Markov jump process model, we translate monthly surveys into minute-scale estimations about behavior and relationships in order to test traditional findings such as meme diffusion and the co-evolution of network topology and node attributes, and to remove errors and fill in missing records in the sensor network data. To prevent overfitting, we set the priors according to patterns that are not controversial in this context, such as the periodicity of human activity, the small-world property of social networks, the 80-20 law of human behavior, and the existence of time and space for events to happen.

Let us assume that two persons became friends over two consecutive monthly-relationship surveys, and both reported convergence in their activities and opinions in behavior surveys. We cannot determine the rich interaction between relationships and behaviors from only the surveys of the coarse temporal resolution. However, by looking at the proximity and location tracking of these two persons, and by relating survey reports to tracking records, we know the order of events in the co-evolution of the social network and the attitudes and behaviors of the individuals – when and where the two persons met initially and became friends, when and where they co-appear afterwards, and how their attitudes and behaviors converge due to interpersonal influence.

Our model of sensor events follows our Bayesian heuristic to locate a person and to anticipate his proximity with other persons when he is out of sight. We form this Bayesian belief of how likely it is that this person will appear at each location, and how likely he is to connect with another person, by weighting where he spends time generally, where he is at a specific time-of-day and day-of-week, where his friends spend time generally and specifically, where people like him often spend time, and how likely it is that he is collocated with other people. In doing this, we enumerate consistent possibilities about location and proximity. For example, two persons cannot be in physical proximity while not being collocated. When we have missing information about location and proximity, we iteratively adjust our Bayesian belief and our imputation of the missing information until we reach maximum likelihood estimation. To this end, we use survey data differently to suit different purposes: when the goal is to achieve the most accurate estimation on missing information about location and proximity, we weight between survey data and previous observations to form a Bayesian belief. When our goal is to find the relationship between survey data and sensor observations, we use survey data as covariates, and fit parameters to achieve maximum likelihood of the sensor data events.

What we have described is a Markov jump process model on the interaction of events and states, and the co-evolution of locations and proximity of a system of individuals. We define the state of this system as the locations of the individuals, and we express the state of C individuals at time t as a state vector: x(t)=(Is person 1 at location 1? Is person 2 at location 2? … Is person C at location L?). The state $x(t)$ of the system is changed by different events $(1, \cdots, v)$, and the state also determines the rates $h_v(x(t))$ at which different events will occur. We use an event vector to describe the number of different events happening in a time window: $r = (r_1, \cdots r_v)'$ where r_i is the number of events of type i. We denote an event by a "reaction" $\sum_i \alpha_i x_i \rightarrow \sum_j \beta_j y_j$, where α_i number of reactant x_i has been consumed and β_j number of product has been generated. In our model of location-proximity co-evolution, individuals change locations either due to their own volition or due to interaction with other individuals, and α_i, β_j are all one.

We are concerned with two types of events in our modeling: change of location not due to interaction with other people, and change of location due to interaction with other people. We express the rate $h_{i:\circ\rightarrow k}(x)$ that an individual i changes location to k due to his own volition as a linear combination of the contributions of different surveyed individual attributes $u_s(i)$, where s indicates different attributes. We express the rate $h_{i:\circ\rightarrow k|j}(x)$ that an individual i changes location to k due to his interaction with individual j who is at location k as a linear combination of surveyed relationships $v_s(i,j)$, where s indicates different pair-wise relations.

$$h_{i:\circ\rightarrow k}(x(t)) = \sum_s b_s(u_s(i), k, t),$$

$$h_{i:\circ\rightarrow k|j}(x(t)) = \sum_s \sum_j 1_{j:k} a_s(v_s(i,j), k, t).$$

In the above, a_s and b_s are parameters. The likelihood of the Markov process is maximized when the rates computed from surveys best fit the rates from the sensors.

We use matrix algebra to express how events change state. To this end we define the reaction matrix A as a $C \times v$ matrix, where C is the length of the state vector $x(t)$ and v is the number of reactions. An element at column k and row j represents the amount added to state $x_j(t)$ if reaction k happens. In our modeling of group dynamics, entries of A are either $+1$ or -1, representing moving into a state (location) or moving out of a state. For example, in the following equation involving four persons and two locations per person, the first three columns of A represent when person 1 moves from location 1 to location 2, person 2 and 3 switch their positions, and speaker 4 moves from location 2 to location 1. The column vector r means an event. If we multiply A by r, we get an update of the state matrix.

$$A \cdot r = \begin{pmatrix} -1 & & & \\ 1 & & & \\ & 1 & & \\ & -1 & \cdots & \\ & -1 & \cdots & \\ & 1 & & \\ & & 1 & \\ & & -1 & \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} \\ 1 \\ -1 \\ -1 \\ 1 \\ \\ \end{pmatrix} = \Delta x$$

Let $r(t_i)$ be the event vector representing the numbers of different events happening between t_i and t_{i+1}. In the Markov jump process, formulation $r_{v_i}(t_i) = 1$ and the other elements of $r(t_i)$ are 0 with probability 1. The system states starting from $x(t_0)$ and corresponding to the sequence of events are updated according to $x(t_{i+1}) = x(t_i) + A \cdot r(t_i)$.

In order to derive the inference algorithm for estimating the dynamics of behavior-interaction co-evolution from noisy sensor data, we begin with the ideal situation that we know all events (t_i, v_i), where $i = 1, \cdots, n$, $0 = t_0 < t_1 < \cdots < t_n = T$, and $v_i \in \{1, \cdots, v\}$. The probability for this sequence of events to occur is

$$P(v, x) = \prod_i h_{v_i}(x_t) \cdot \exp\left(-\sum_i h_{v_i}(x_t) \cdot (t_{i+1} - t_i)\right)$$

In reality, we have only discrete time noisy observations $y(n \cdot \Delta t)$ regarding Wi-Fi hotspot accessibility and proximity, and we want to infer from these discrete time observations how many, when, and what events happened between these observations. The inference algorithm becomes non-trivial when the time interval between two consecutive observations becomes large, when we have missing data, and when we have data that are incompatible with the model. However it is possible to construct exact MCMC algorithms for inference based on discrete time observations, and it is possible to posit inferences through mean field approximation and variational method [22].

We introduced the following approximations to make the inference of co-evolution dynamics much simpler. Our first approximation is that events occur only at the times of observation, and this approximation introduces 12-minute error into event times. Our second approximation is that the observations for inferring the state have joint Gaussian distributions conditioned on that state.

Thus the probability of a sequence of latent events $v(t)$, together with the corresponding latent states $x(t)$ and observations $y(t)$, is as follows:

$$P(v, x, y) = \prod_{t,c} P\left(y^{(c)}(t) \middle| x^{(c)}(t)\right) \cdot P(x(0)) \prod_t P\left(v(x(t))\right),$$

$$P(v(x(t)) = 0) = \exp\left(-\sum_j h_j(x_t)\right),$$

$$P(v(x(t)) = i) = \frac{h_i(x(t))}{\sum_j h_j(x(t))}\left(1 - \exp\left(-\sum_j h_j(x(t))\right)\right),$$

$$P\left(y^{(c)}(t) \middle| x^{(c)}(t)\right) = N\left(\mu_{x^{(c)}(t)}, \Sigma_{x^{(c)}(t)}\right).$$

We use Gibbs sampling to infer latent states and parameters:

$$v(x(t)) | \mu, \Sigma \sim \text{Categorical}\left(\left\{P(v(x(t)) = i) \cdot P(y(t) | x(t)) P\left(v(x(t+1))\right) : i = 0, \cdots, v\right\}\right),$$

$$h_i \sim \text{Dirichlet}\left(h + \sum_t \delta_{v(x(t)),i}\right),$$

$$\Sigma | \kappa_0, \mu_0, v_0, \Psi_0 \sim W^{-1}\left(T + \kappa_0, \Psi_0 + \sum_t (y_t - \bar{y})(y_t - \bar{y})^T + \frac{n v_0}{n + v_0}(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T\right),$$

$$\mu | \Sigma, v_0, \mu_0 \sim N\left(\frac{n\bar{x} + v_0 \mu_0}{T + v_0}, \frac{\Sigma}{n + v_0}\right).$$

The Markov jump process model is related to the influence model [1]. In the Markov jump model, system state is updated by multiplying a reaction matrix and an event vector. In the influence model, system state is updated by multiplying an influence matrix and a marginal state vector. The two models can be translated into one another by translating between the event vector and the marginal state vector and subsequently translating between the reaction matrix and influence matrix.

The Markov jump process model is also related to the temporal exponential random graph model (tERGM) [10] in that both models can encode different statistics about network topology in the likelihood function, and can subsequently study the evolution of network topology. However, tERGM requires that the network topology at each time step be in equilibrium in terms of the concerned network topology statistics, and this requirement make tERGM inappropriate for studying dynamics in networks where equilibrium is not attained.

We can gain insight into this Markov jump process model from a simplification of the model. In this simplification, the surveyed relationships and attributes are constant (that is, time is homogeneous) within each survey period, and the rate that a location or a proximity instance is detected is also constant within each survey period. Without confirmed reports about the probability distributions of sensor data regarding the average behavior, we assume normal distribution, and focus on the average behavior. This simplification results in linear regression models. The rate that a person is detected at a location is a linear combination of the surveyed attributes about the person, and $b_k^{(s)}$ represents how a surveyed attribute contributes to the rate of location detection. The rate that a dyad between two persons is detected is a linear combination of the surveyed relations about this dyad, and $a_k^{(s)}$ represents how a surveyed relationship contributes to the rate of dyad detection. By relating the rate of location detection and the rate of proximity detection $r_p = f(r_k) + \text{error}$, we have built a generalized regression model to explain a surveyed relationship with surveyed locations $\sum_s a_k^{(s)} y_{ij}^{(s)}(t) = f\left(\sum_s b_k^{(s)} x_i^{(s)}(t)\right) + \text{error}$, or to explain a surveyed relationship with location and proximity sensing, and so on.

The potential of a Markov model of human dynamics is supported by the previous successful applications of regression models to predict relationships from sensor data. Eagle, for example, reported that sensor data such as off-campus proximity during weekends and nights, who called whom, and who sent short messages to whom could explain friendship with 95% accuracy [2].

A Markov process model gives us more power than a regression model when sample data are scarce, or data have missing or erroneous content. In the task of identifying the dyads in a cluster of friends from a short time window of observation, we have a small chance of observing calls or co-occurrence between all pairs, but we have a larger chance of observing higher rates of a potential calls or co-occurrence within this cluster.

5. EXPERIMENTAL RESULTS

We demonstrate that relationships and individual behaviors interact with each other. In our analysis, we propose a dynamical process model to track the co-evolution of relationships and

behavior. We can often relate sensor data (where people go and who is around them) to labels such as friendship, number of years in school, and activities, by applying machine-learning methods. In this section, we put the structure of human dynamics and its modeling into useful applications.

5.1 Predicting relations

Our first application tracks friendship based on how friendship is related to factors such as where people live, how many years they have been in school, where people go, and how often people go together. Tracking friendship is useful, because friendship shapes many of the aspects of behavior that we previously described, or that other researchers have documented, and tracking friendship through surveys is neither user-friendly nor indicative of friendship diagnostics. There will be no difficulty in using the same technique to track behaviors and attributes such as happiness and fitness, given that a relation between sensor data and attributes exists.

Prior to the availability of sensor data for understanding human behavior, researchers predicted friendship based on the collected labels about individuals and some known relationships. Many researchers adopted the model that says whether A and B are friends is positively related to whether A or B likes to make many friends (sociability), whether A and B have many characteristics in common (assortativity), and how many common friends A and B have (triangle closure). They proceeded to use logistic regression to relate the log odds of friendship to a linear combination of sociability, assortativity, and triangle closure. Logistic regression and surveyed attributes can predict friendship well in the Social Evolution data, because people in the same dormitory sector and people in the same years in school are five times more likely to be friends. Logistic regression and surveyed attributes can even predict how new links were added into the friendship network of the residents if we add the interaction between the elapsed time since the beginning of a semester and the students' years in school, because freshmen made new friends over time and the friendship relations of other students were stable.

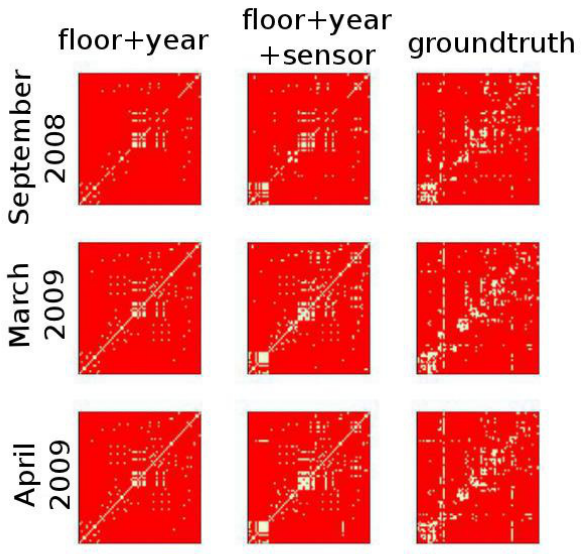

Figure 9: Dormitory sector, year in school, and time since enrollment as a freshman explain the structure and evolution of friendship (left column). Proximity and place-visiting similarity explain more variations about friendship (middle column).

However, the prediction based on the traditional data collection method fails to predict the differences among individuals who have the same predictor attributes (in the Social Evolution data set, dormitory sector and years in school) and how friendship varies daily, which are perhaps more useful for the residents. The sensor data solves this problem by allowing us to model the everyday locations and proximities collected by sensors, which are probabilistically conditioned on the attributes collected by surveys. In other words, the sensor data now sit between the relationship to predictors and the surveyed attributes.

To evaluate sensor data in estimating friendship, we simulate our dynamical process model conditioned on the survey data to find the log-odds that two persons were friends, and compare the log-odds and the explained deviance with those from the logistic regression model. We constructed the logistic regression model to predict, from whether two persons share the same dormitory sector, whether they are the same year in school, and how long the experiment had been conducted if one person was a freshman, whether A reported B as a friend in the surveys. ANOVA shows that a shared dormitory sector explains about 8% variance, shared year explains an additional 4%, the evolution of a freshmen's friendship links explains 4%, and the sensor data explains an additional 6%. Overall 22% variance of friendship relation is explained, where friendship relation has binomial distribution.

Figure 9 shows that by using the surveyed attributes of individuals and logistic regression we can correctly identify the structure of the friendship network of the student dorm residents, and using the sensor data (where people were, how often people were together) and the simulation approach we can not only capture the structure induced by common dorm sector and year, but also verify whether a relationship really exists when we predict that it should.

It could be misleading to use precision (the fraction of retrieved instances that are relevant) and recall (the fraction of relevant instances that are retrieved) to access a method to detect friendship, and a better way is to ask whether the method correctly captures the structure statistics of friendship. In the Social Evolution data, a person on average had five to six friends. Only 5% of all pairs were friendship pairs. Hence, a null classifier could easily archive 95% accuracy by saying that all pairs are non-friend pairs. Both logistic regression and simulation-based approaches identify the structure of the friendship network.

5.2 Predicting sensor data from the survey

Our second application samples typical time series regarding where people go and how people come together based on the trained dynamical process model. Sampling time series is important for protecting privacy and for planning and evaluating intervention.

The privacy issues arise because the behavioral data collected by sensors have considerable commercial and research value, and collecting and distributing such data often involves significant legal and ethnical issues. The Social Evolution data, for example, cannot leave a dedicated server. Distributing synthesized sensor data has good potential in addressing the privacy issue, because by fusing sensor data from many subjects of the same attributes into synthesized typical behavior and interaction, we can both get rid of subject-identifiable features and also control the level of detail that we provide. Synthesizing time series also enables us to simulate

different ways of shaping group or individual behavior before we put such methods into practice. For example, if we put an attraction at the athletics center, how likely is this attraction to be picked up by an individual in the student dorm, how likely is it to attract the individual again and to foster regular aerobic exercise, and how likely will the influenced individual be in turn to influence other individuals?

While the mathematics of the Markov jump process is complicated, simulating behavior and interaction is straightforward. At any given time, an individual decides whether to leave his current place. If he does not leave his current place, he will decide whether to leave a moment later. If he leaves his current place, he then decides which next place to go to, or which friend to go to. If he has seen a non-friend a lot in the previous two weeks, he then decides whether he should make friends with that person. If he hasn't seen a friend in the previous two weeks, he then decides whether he should turn this person into a non-friend. The rates at which he makes different decisions can be found from persons like him (that is, same year, living in the same dorm sector, or working in the same department) in a training data set.

Figure 10 illustrates two synthesized paths. One is synthesized from a computer science senior who has friends working in the media laboratory and who undertakes regular physical exercise. Another path is synthesized from a biology sophomore who is quite arduous in physical exercise. Such paths are typical of how a computer science senior or a biology sophomore moves around every day. However, such paths do not exist in real life.

6. CONCLUSION AND FUTURE WORKS

How social relationships and individual behaviors co-evolve in time and space has important implications, but is poorly understood due to the lack of data. We have shown evidence that relationships and behaviors co-evolve in a student dormitory, based on monthly surveys and locations/proximities tracked by cell phones for a nine-months period. We describe a Markov jump process model to capture this co-evolution in terms of the rates of going to places and friends. We demonstrate that by modeling the dynamics in sensor data, we can predict friendship, and can synthesize useful and accurate behavior and interaction projections.

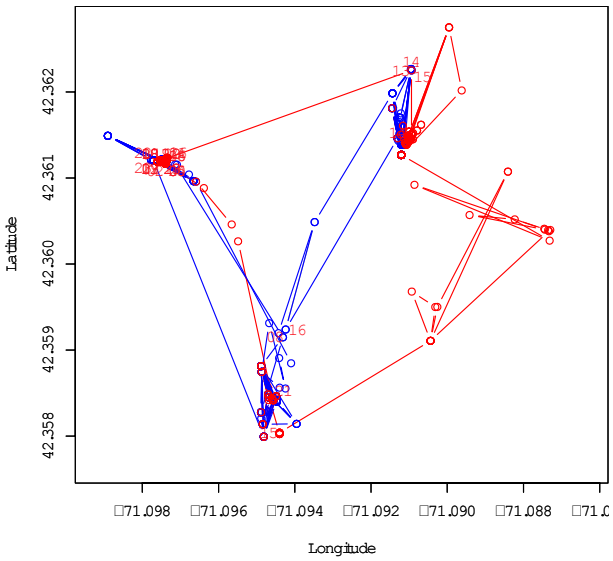

Figure 10: The typical day of a computer science senior (red) and of a biology sophomore (blue). Red numbers represent the hour of day.

7. ACKNOWLEDGEMENTS

Research was sponsored by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, and by AFOSR under Award Number FA9550-10-1-0122. Views and conclusions in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation.

Bruno Lepri's research is funded by PERSI project inside the Marie Curie Cofund 7th Framework.

8. REFERENCES

[1] Asavarithiratham, C., Roy, S., Lesieutre, B., and Verghese, G. (2001). The Influence Model. IEEE Control Systems Magazine. 21, 6, pp. 52-64.

[2] Cho, E., Myers, S., and Leskovec, J. (2011). Friendship and Mobility: User Movement In Location-Based Social Networks. In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'11), pp. 1082-1090.

[3] Christakis, N. A., Fowler, J. H. (2007). The Spread of Obesity in a Large Social Network over 32 Years. The New England Journal of Medicine. 357, pp. 370–379.

[4] Christakis, N. A., Fowler, J. H. (2008). The Collective Dynamics of Smoking in a Large Social Network. The New England Journal of Medicine. 358 (May 2008), pp 2249–2258.

[5] Eagle, N., Pentland, A., and Lazer, D. (2009) Inferring Friendship Network Structure by using Mobile Phone Data. In Proceedings of the National Academy of Sciences of the United States of America. 106, 36, pp. 15274-15278.

[6] Fowler, J., and Christakis, N., (2008). Dynamic Spread of Happiness in a Large Social Network: longitudinal analysis over 20 years in the Framingham Heart Study, BMJ, vol. 337, p. a2338.

[7] Frank, O., and Strauss, D., (1986). Markov graphs. Journal of the American Statistical Association 81, 832–842.

[8] Gonzalez, A., Hidalgo, C., and Barabasi, A-L. (2008) Understanding individual human mobility patterns. Nature 453(5) June 2008

[9] Guo, F., Hanneke, S., Fu, W., and Xing, E. P. (2007). Recovering temporally rewiring networks: a model-based approach. In Proceedings of the 24th International Conference on Machine Learning. ICML'07, pp. 321–328.

[10] Hanneke S., Fu, W. and Xing, E.P. (2010) Discrete temporal models of social networks. In Electronic Journal of Statistics. Vol. 4, pp. 585-605

[11] Horvitz, E., Apacible, J. Sarin, R., and Liao, L. (2005). Prediction, Expectation, and Surprise: Methods, Designs, and Study of a Deployed Traffic Forecasting Service. In Proceedings of Twenty-First Conference on Uncertainty in Artificial Intelligence, (UAI-2005).

[12] Kossinets G., and Watts, D. (2006) Empirical Analysis of an Evolving Social Network. Science, 311 (5757), pp. 88-90.

[13] Leenders, R.Th.A.J. (1997) Longitudinal behavior of network structures and actor attributes: modeling interdependcnce of contagion and selection. In P. Doreian and F.N. Stokman (eds.) Evolution of social networks (pp. 165-184). Amsterdam: Gordon and Breach. 1997

[14] Lyons, Russell (2011) The Spread of Evidence-Poor Medicine via Flawed Social-Network Analysis. Statistics, Politics, and Policy: Vol. 2: Iss. 1, Article 2.

[15] Madan, A., Cebrian, M., Lazer, D., and Pentland, A. (2010) Social Sensing for Epidemiological Behavior Change. In Proceedings of 12th ACM International Conference on Ubiquitous Computing, pp. 291-300

[16] Madan A., Farrahi K., Daniel Gatica-Perez and Pentland A. (2009) Pervasive Sensing to Model Political Opinions in Face-to-Face Networks, In Proceedings of Pervasive 2011.

[17] Madan A., Moturu, S., Lazer, D., and Pentland, A. (2010) Social Sensing: Obesity, Unhealthy Eating and Exercise in Face-to-Face Networks. In Wireless Health 2010 pp. 104-110.

[18] Madan, A., and Pentland, A. (2009) Modeling Social Diffusion Phenomena Using Reality Mining. In AAAI Spring Symposium on Human Behavior Modeling, 2009.

[19] Noulas, A., Musolesi, M., Pontil M., and Mascolo, C. (2009) "Inferring Interests from Mobility and Social Interactions". In Proceedings of Neural Information Processing Systems (NIPS) 2009

[20] Olguín, D., Waber, B., Kim, T., Mohan, A., Ara, A., and Pentland, A. (2009) Sensible Organizations: Technology and Methodology for Automatically Measuring Organizational Behavior. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics. Vol. 39, No. 1, pp. 43-55. February, 2009

[21] Robins, G.L., Pattison, P., Kalish, Y., and Lusher, D. An Introduction to Exponential Random Graph (p*) Models for Social Networks. Social Networks.

[22] Wainwright, M.J., and Jordan, M.I. (2008). Graphical Models, Exponential Families, and Variational Inference. Foundations and Trends in Machine Learning. Vol. 1. N° 1-2, pp. 1-305

[23] Wasserman, S. and Pattison, P. (1996). Logit models and logistic regression for social networks. I. An introduction to Markov graphs and p*. Psychometrika, 61, pp. 401-425.

[24] Wren, C.R., Ivanov, Y.A., Leigh, D., Westhues, J. (2007). The MERL Motion Detector Dataset. In ICMI Workshop on Massive Datasets (MD), pp. 10-14,