ORIGINAL ARTICLE

Nathan Eagle · Alex (Sandy) Pentland

Reality mining: sensing complex social systems

Abstract We introduce a system for sensing complex social systems with data collected from 100 mobile phones over the course of 9 months. We demonstrate the ability to use standard Bluetooth-enabled mobile telephones to measure information access and use in different contexts, recognize social patterns in daily user activity, infer relationships, identify socially significant locations, and model organizational rhythms.

Keywords Mobile phones · Bluetooth · Complex social systems · Wearable computing · User modeling

1 Introduction

The last 10 years could rightly be coined the decade of the mobile phone. In 2004, over 600 million handsets were sold, dwarfing the number of personal computers sold that year [27]. The potential functionality of this ubiquitous infrastructure of mobile devices is dramatically increasing. In this paper we describe how data collected from mobile phones can be used to uncover regular rules and structure in the behavior of both individuals and organizations. In Sect. 2, we begin with a discussion of the rationale for using phones as wearable sensors and the type of data they can collect. Subsequently, Sect. 3 describes the benefits of modeling individual users by fusing information from cell towers with discovered Bluetooth IDs. Turning our attention away from individuals and towards dyads, in Sect. 4 we extract salient features indicative of the relationships between subjects using proximity, time, and location data. Finally, with the nodes and edges of this social network identified, the concept of organizational rhythms is

introduced as a useful metric for quantifying organizational behavior.

## 2 Mobile phones as wearable sensors

For over a century, social scientists have conducted surveys to learn about human behavior. However, surveys are plagued with issues such as bias, sparsity of data, and lack of continuity between discrete questionnaires. It is this absence of dense, continuous data that also hinders the machine learning and agent-based modeling communities from constructing more comprehensive predictive models of human dynamics. Over the last two decades there has been a significant amount of research attempting to address these issues by building location-aware devices capable of collecting rich behavioral data [1, 6, 11, 16, 22, 24].

Although these projects were relatively successful, by depending on a limited supply of custom hardware, they were unsuitable for groups of a greater size. While drawing extensively on previous work from the Ubiquitous Computing field, one of the contributions of this paper is to show the potential for these ideas to scale upwards. With the rapid technology adoption of mobile phones comes an opportunity to collect a much larger dataset on human behavior [10, 18]. The very nature of mobile phones makes them an ideal vehicle to study both individuals and organizations: people habitually carry their mobile phones with them and use them as a medium for much of their communication. In this paper we capture all the information to which the phone has access (with the exception of content from phone calls or text messages) and describe how it can be used to provide insight into both the individual and the collective.

2.1 Mobile phone proximity logs

One of the key ideas in this paper is to exploit the fact that modern phones use both a short-range RF network

N. Eagle (✉) · A. (Sandy) Pentland
MIT Media Laboratory, 20 Ames Street,
Cambridge, MA, 02139 USA
E-mail: nathan@media.mit.edu
E-mail: sandy@media.mit.edu
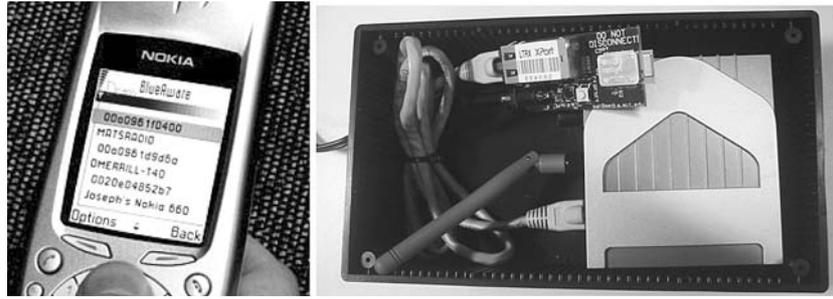
Fig. 1 Methods of detecting Bluetooth devices—BlueAware and Bluedar. BlueAware (*left*) is running in the foreground on a Nokia 3650. BlueAware is an application that runs on Symbian Series 60 phones. It runs in the background and performs repeated Bluetooth scans of the environment every 5 min. Bluedar (*right*) is comprised of a Bluetooth beacon coupled with a WiFi bridge. It also performs cyclic Bluetooth scans and sends the resulting BTIDs over the 802.11b network to the Reality Mining server

(e.g., Bluetooth) and a long-range RF network (e.g., GSM), and that the two networks can augment each other for location and activity inference. The idea of logging cell tower ID to determine approximate location will be familiar to readers, but the idea of logging Bluetooth devices is relatively recent and provides different types of information [15].

Bluetooth is a wireless protocol in the 2.40–2.48 GHz range, developed by Ericsson in 1994 and released in 1998 as a serial-cable replacement to connect different devices. Although market adoption has been initially slow, according to industry research estimates by 2006 90% of PDAs, 80% of laptops, and 75% of mobile phones will be shipped with Bluetooth [28]. Every Bluetooth device is capable of "device-discovery," which allows them to collect information on other Bluetooth devices within 5–10 m. This information includes the Bluetooth MAC address (BTID), device name, and device type. The BTID is a 12-digit hex number unique to the particular device. The device name can be set at the user's discretion; e.g., "Tony's Nokia." Finally, the device type is a set of three integers that correspond to the device discovered; e.g., Nokia mobile phone or IBM laptop.

To log BTIDs we designed a software application, BlueAware, that runs passively in the background on MIDP2-enabled mobile phones. Bluetooth was primarily designed to enable wireless headsets or laptops to connect to phones, but as a by-product devices are becoming aware of other Bluetooth devices carried by people nearby. Our application records and timestamps the BTIDs encountered in a proximity log and makes them available to other applications, similar to the Jabberwocky project developed by Paulos et al. [19]. BlueAware is automatically run in the background when the phone is turned on, making it essentially invisible to the user.

A variation on BlueAware is Bluedar. Bluedar, shown in Fig. 1 (right), was developed to be placed in a social setting and continuously scan for visible devices, wirelessly transmitting detected BTIDs to a server over an 802.11b network. The heart of the device is a Bluetooth beacon designed by Mat Laibowitz, incorporating a class 2 Bluetooth chipset that can be controlled by an XPort web server [14]. We integrated this beacon with an 802.11b wireless bridge and packaged them in an unobtrusive box. An application was written to continuously telnet into multiple BlueDar systems, repeatedly scan for Bluetooth devices, and transmit the discovered proximate BTIDs to our server. Because the Bluetooth chipset is a class 2 device, it is able to detect any visible Bluetooth device within a working range of up to 25 m. We are currently using the system to prototype a proximity-based introduction service [9].

2.1.1 Refresh rate versus battery-life

Continually scanning and logging BTIDs can expend an older mobile phone battery in about 18 h[1]. While continuous scans provide a rich depiction of a user's dynamic environment, most individuals expect phones to have standby times exceeding 48 h. Therefore BlueAware was modified to only scan the environment once every 5 min, providing at least 36 h of standby time.

2.2 Privacy implications

Mining the reality of our 100 users raises justifiable concerns over privacy. However, the work in this paper is a social science experiment, conducted with human subject approval and consent of the users. Outside the lab we envision a future where phones will have greater computational power and will be able to make relevant inferences using only data available to the user's phone. In this future scenario, the inferences are done in real-time on the local device, making it unnecessary for private information to be taken off the handset. However, the computational models we are currently using cannot be implemented on today's phones. Thus, our results aim to show the potential of the information that can be gleaned from the phone, rather than presenting a system

---

[1]Using a 6-month old battery of a Nokia 6600 in a sparsely populated Bluetooth environment.

that can be deployed today outside the realm of research.

2.3 The dataset

Our study consists of 100 Nokia 6600 smart phones pre-installed with several pieces of software we have developed as well as a version of the Context application from the University of Helsinki [20]. Seventy-five users are either students or faculty in the MIT Media Laboratory, while the remaining twenty-five are incoming students at the MIT Sloan business school adjacent to the Laboratory. Of the 75 users at the Lab, 20 are incoming master's students and 5 are incoming MIT freshman. The information we are collecting includes call logs, Bluetooth devices in proximity, cell tower IDs, application usage, and phone status (such as charging and idle), and comes primarily from the Context application. The study has generated data collected by 100 human subjects over the course of the academic year that represent approximately 450,000 h of information about users' location, communication and device usage behavior. We released a public, anonymous version of the dataset on our website http://reality.media.mit.edu.

3 User modeling: identifying structure in routine

Although humans have the potential for relatively random patterns of behavior, there are easily identifiable routines in every person's life. These can be found on a range of timescales: from the daily routines of getting out of bed, eating lunch, and driving home from work, to weekly patterns such as Saturday afternoon softball games, to yearly patterns like seeing family during the holidays in December. While our ultimate goal is to create a predictive classifier that can perceive aspects of a user's life more accurately than a human observer (including the actual user), we begin by building simple mechanisms that can recognize many of the common structures in the user's routine. Learning the structure of an individual's routine has already been demonstrated using other modalities, however we present this analysis as a foundation which will then be extended to demonstrate the learning of *social* structures.

We begin with a simple model of behavior in three states: home, work, and elsewhere. The data are obtained from Bluetooth, cell tower, and temporal information collected from the phones. We then incorporate information from static Bluetooth devices (such as desktop computers), using them as 'cell towers' to identify significant locations and localize the user to a ten-meter radius. We show that most users spend a significant amount of time in the presence of static Bluetooth devices, particularly when they don't have cell tower reception (e.g., inside the office building). This makes them an ideal supplement to cell towers for location classification.

3.1 Location based on cell towers and Bluetooth

There has been a significant amount of research which correlates cell tower ID with a user's location [3, 4, 12]. For example, Laasonen et al. [13] describe a method of inferring significant locations from cell tower information through analysis of the adjacency matrix formed by proximate towers. They were able to show reasonable route recognition rates, and most importantly, succeeded in running their algorithms directly on the mobile phone.

Obtaining accurate location information from cell towers is complicated by the fact that phones can detect cell towers that are several miles away. Furthermore, in urban areas it is not uncommon to be within range of several dozen different towers. The inclusion of information on all the current visible towers as well as their respective signal strengths would help solve the location classification problem, although multipath distortion may still confound estimates.

We observe that relatively high location accuracy may also be achieved if the user spends enough time in one place to provide an estimate of the location's cell tower *probability density function*. A phone in some static location is associated with different cell towers at different times. Thus, it is possible to generate the distribution of time spent associated with a set of towers for a particular area. This distribution of detected towers can vary substantially with even small changes in location. Figure 2 shows the distribution of cell towers seen for a given area with a 10 m radius. Towers were only included in these distributions if the common area's static Bluetooth desktop computer was also visible, ensuring the users' location within 10 m (or less). Discrepancies in the distributions are attributed to the users' typical position within the 10 m radius. Users 4 and 5 both share a window office and have virtually the same cell tower distribution, despite having a very different distribution of hours spent in the office (as verified by the Bluetooth and cell tower logs). Users 1 and 2 both spend the majority of their time in the common area away from the windows and see only half as many towers as the others. User 3 is in a second office in the same area, and has a distribution of cell towers that is intermediate between the two other sets of users.

Despite progress in mapping a cell tower to a location, the resolution simply cannot be as high as many location-based services require. GPS is an alternative approach that has been used for location detection and classification [2, 17, 25] but the line-of-sight requirements prohibit it from working well indoors. We have therefore incorporated the use of static Bluetooth device ID as an additional indicator of location, and shown that it provides a significant improvement in user localization, especially within office environments. This fusion of data is particularly appropriate since areas where cellular signals are weak, such as in the middle of large buildings, often correspond to places where there are many static Bluetooth devices, such as desktop

Cell Tower Distributions for a 10m radius
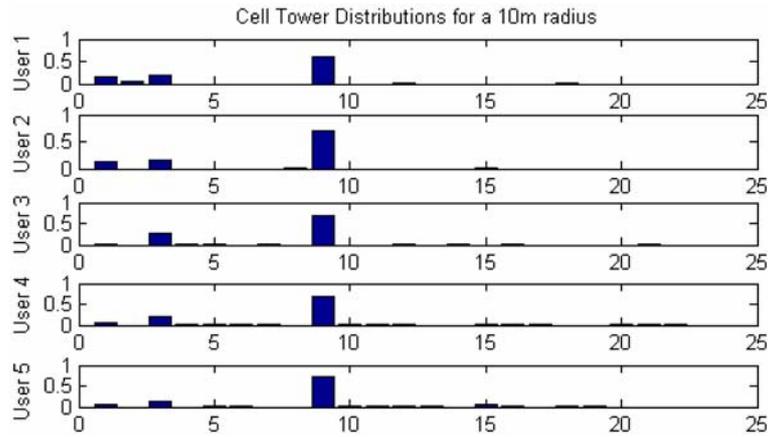
Fig. 2 Cell tower probability density functions. The probability of being associated with one of the 25 visible cell towers is plotted above for five users who work on the third floor corner of the same office building. Each tower is listed on the *x*-axis and the probability of the phone logging it while the user is in his office is shown on the *y*-axis. (Range was assured to 10 m by the presence

of a static Bluetooth device.) It can be seen that each user 'sees' a different distribution of cell towers depending on the location of his office, with the exception of Users 4 and 5, who are officemates and have the same distribution despite being in the office at different times

computers. On average, the subjects in our study were without mobile phone reception for 6% of the time. When they did not have reception, however, they were within the range of a static Bluetooth device or another mobile phone 21% and 29% of their time, respectively. We expect the coverage by Bluetooth devices to increase dramatically in the near future as they become more common in computers and electronic equipment.

If this trend continues, Bluetooth IDs may become as important as cell tower mapping for estimation of user location. Figure 3 shows the 10 most frequently detected Bluetooth devices for one subject averaged for the month of January. This figure not only provides insight into the times the user is in his office (from the frequencies of the 'Desktop'), but, as mentioned in Sect. 4, also into the type of relationship with other subjects. For example, the figure suggests the user leaves his office during the hour of 14:00 and becomes increasingly

proximate to Subject 4. Judging from the strong cutoffs at 9:00 and 17:00, it is clear that this subject had very regular hours during the month, and thus has fairly predictable high-level behavior. This "low entropy" behavior is also depicted in Fig. 4.

3.2 Models to identify location and activity

Human life is inherently imbued with routine across all temporal scales, from minute-to-minute actions to monthly or yearly patterns. Many of these patterns in behavior are easy to recognize, however some are more subtle. We attempt to quantify the amount of predictable structure in an individual's life using an information entropy metric. In information theory, the amount of randomness in a signal corresponds to its entropy, as defined in 1938 by Claude Shannon in the equation below.
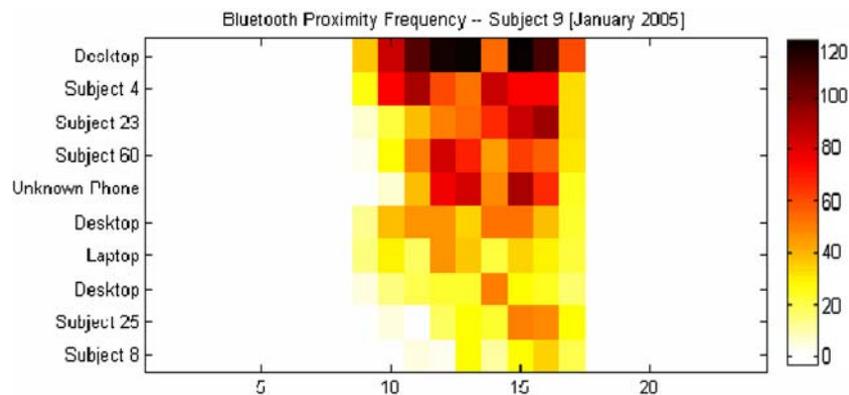
Fig. 3 The top 10 Bluetooth devices encountered for Subject 9 during the month of January. The subject is only regularly proximate to other Bluetooth devices between 9:00 and 17:00, while at work—but never at any other times. This predictable

behavior will be defined in Chap. 4 as 'low entropy.' The subject's desktop computer is logged most frequently throughout the day, with the exception of the hour between 14:00 and 15:00. During this time window, Subject 9 is most often proximate to Subject 4
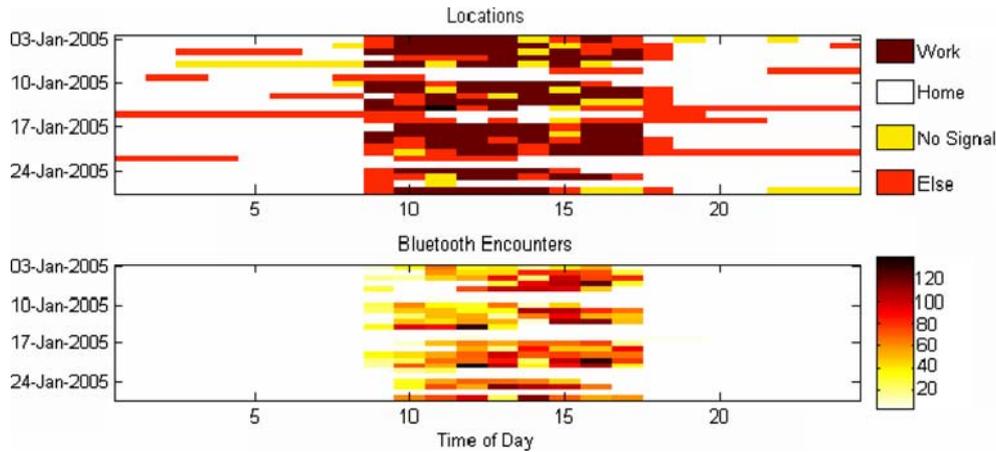
Fig. 4 A 'low-entropy' ($H = 30.9$) subject's daily distribution of home/work transitions and Bluetooth devices encounters during the month of January. The top figure shows the most likely location of the subject: "Work, Home, Elsewhere, and No Signal." While the subject's state sporadically jumps to "No Signal," the other states occur with very regular frequency. This is confirmed by the Bluetooth encounters plotted below representing the structured working schedule of the 'low-entropy' subject

$$H(x) = -\sum_{i=1}^{n} p(i) \log_2 p(i).$$

For a more concrete example, consider the problem of image compression (such as the jpeg standard) of an overhead photo taken of just an empty checkerboard. This image (in theory) can be significantly compressed because it does not contain much 'information'. Essentially the entire image could be recreated with the same, simple pattern. However, if the picture was taken during the middle of a match, the pieces on the board introduce more randomness into the image and therefore it will prove to be a larger file because it contains more information, or entropy.

Similarly, people who live entropic lives tend to be more variable and harder to predict, while low-entropy lives are characterized by strong patterns across all time scales. Figure 4 depicts the patterns in cell tower transitions and the total number of Bluetooth devices encountered at each hour during the month of January for Subject 9, a 'low-entropy' subject.

It is clear that the subject is typically at home during the evening and all night until 8:00, when he commutes in to work, and then stays at work until 17:00 when he returns home. We can see that almost all of the Bluetooth devices are detected during these regular office hours, Monday through Friday. This is certainly not the case for many of the subjects. Figure 5 displays a different set of behaviors for Subject 8. The subject has much less regular patterns of location and in the evenings has other mobile devices in close proximity. We will use contextualized information about proximity with other mobile devices to infer relationships, described in Sect. 4.

While calculating a life's entropy be used as a method of self-reflection on the routines (or ruts) in one's life, it can also be used to compare the behaviors of different demographics. Figure 6 shows the average weekly entropy of each of the demographics in our study, based on their location {work, home, no signal, elsewhere} each hour. Average weekly entropy was calculated by drawing 100 samples of a 7-day period

Fig. 5 A 'high entropy' ($H = 48.5$) subject's daily distribution of home/work transitions and Bluetooth device encounters during the month of January. In contrast to Fig. 4, the lack of readily apparently routine and structure makes this subject's behavior harder to model and predict
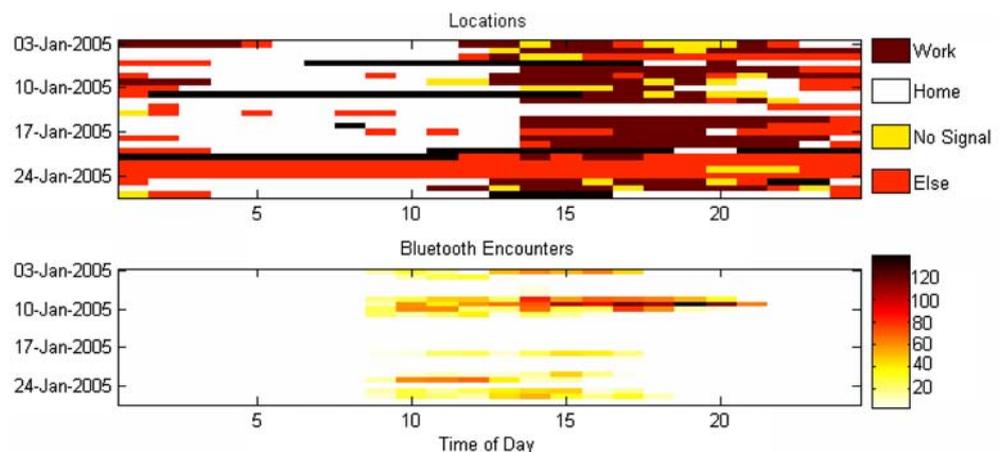
Fig. 6 Entropy, $H(x)$, was calculated from the {work, home, no signal, elsewhere} set of behaviors for 100 samples of a 7-day period. The Media Lab freshmen have the least predictable schedules, which makes sense because they come to the lab much less regular basis. The staff and faculty have the most least entropic schedules, typically adhering to a consistent work routine
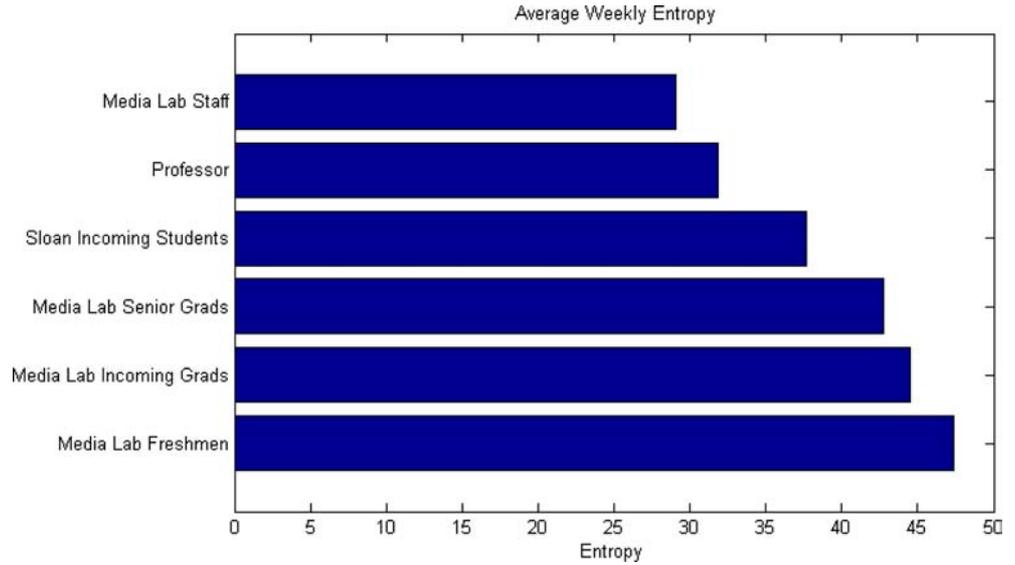

for each subject in the study. No surprise to most, the Media Lab first-year undergraduates are the most entropic of the group. The freshmen do not come into the lab on a regular basis and have seemingly random behavior with $H(x) = 47.3$ (the entropy of a sequence of 168 random numbers is approximately 60). The graduate students (Media Lab incoming, Media Lab senior, and Sloan incoming) are the next most entropic with $H(x) = \{44.5, 42.8, 37.6\}$ respectively. Finally, the Media Lab faculty and staff have most rigidity in their schedules, reflected in their relatively low-average entropy measures, $H(x) = \{31.8, 29.1\}$.

One similarity between the different demographics shown above is the clear role time plays in determining user behavior. To account for this, we have developed a simple Hidden Markov Model, shown in Fig. 7, conditioned on both the hour of day $\left(T^1 \in \{1, 2, 3..., 24\}\right)$ as well as on weekday or weekend $\left(T^2 \in \{1, 2\}\right)$. Initially observations in the model are simply the distri-

bution of cell towers $\left(Y^1 \in \{CT_1, CT_1, ..., CT_{n_1}\}\right)$ and Bluetooth devices $\left(Y^2 \in \{BT_1, BT_1, ..., BT_{n_2}\}\right)$. A straightforward Expectation-Maximization inference engine was used to learn the parameters in the transition model, $P(Q_t|Q_{t-1})$, and the observation model $P(Y_t|Q_t)$, and performed clustering in which we defined the dimensionality of the state space. The hidden state is represented in terms of a single discrete random variable corresponding to three different situations, $Q \in \{$home, work, other$\}$. After training our model with one month of data from several subjects we were able to provide a good separation of clusters, typically with greater than 95% accuracy. Examination of the data shows that non-linear techniques will be required to obtain significantly higher accuracy. However, for the purposes of this chapter, this accuracy has proven sufficient. In future work we hope to leverage the information within LifeNet [23] to create more specific inferences about activity.

Fig. 7 A Hidden Markov Model conditioned on time for situation identification. The model was designed to be able to incorporate many additional observation vectors such as friends nearby, traveling, sleeping and talking on the phone
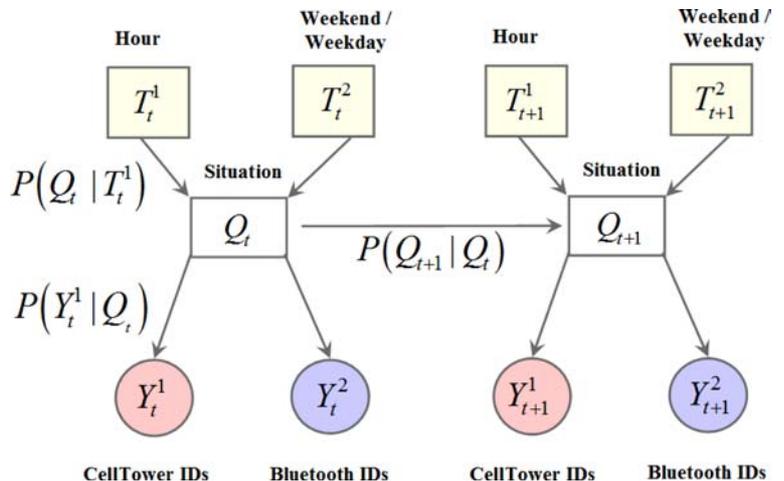
3.3 Mobile usage patterns in context

Capturing mobile phone usage patterns of 100 people for an extended period of time can provide insight into both the users and the ease of use of the device itself. For example, 35% of our subjects use the clock application on a regular basis (primarily to set the alarm clock and then subsequently to press snooze), yet it takes 10 keystrokes to open the application from the phone's default settings. Not surprisingly, specific applications, such as the alarm clock, seem to be used much more often at home than at work. Figure 8 is a graph of the aggregate popularity of different applications when both at home and at work. It is interesting to note that despite the subjects being technically savvy, there was not a significant amount of usage in the sophisticated features of the phone—indeed the default game "Snake" was used just as much as the elaborate Media Player application.

While there is much to be gained from a contextual analysis of application usage, perhaps the most important and still most popular use of the mobile phone is as a communication device. Figure 9 is a breakdown of the different types of usage patterns from a selection of the subjects. Approximately 81% of communication on the phone was completed by placing or receiving a voice call. Data (primarily email) was at 13% of the communication, while text messaging was 5%.

Learning a user's application routines can enable the phone to place a well-used application in more prominent places, for example, as well as creating a better model of the behavior of an individual [26]. As we shall see in Sect. 4, these models can also be augmented with additional information about a user's social context.

3.4 Data characterization and validation

This section describes how errors may be introduced into the data through data corruption, device detection failures, and most significantly, through human error (Fig. 10).

3.4.1 Data corruption

All the data from a phone are stored on a flash memory card, which has a finite number of read–write cycles. Initial versions of our application wrote over the same cells of the memory card. This led to the failure of a new card after about a month of data collection, resulting in the complete loss of data. When the application was changed to store the incremental logs in RAM and subsequently write each complete log to the flash memory, our data corruption issues virtually vanished. However, 10 cards were lost before this problem was

Fig. 8 Average application usage in three locations (other, work, and home) for 100 subjects. The x-axis displays the fraction of time each application is used, as a function of total application usage. For example, the usage at home of the clock application comprises almost 3% of the total times the phone is used. The 'phone' application itself comprises more than 80% of the total usage and was not included in this figure
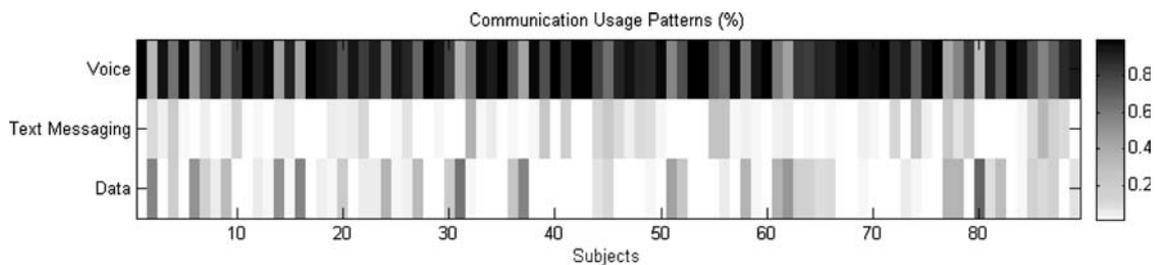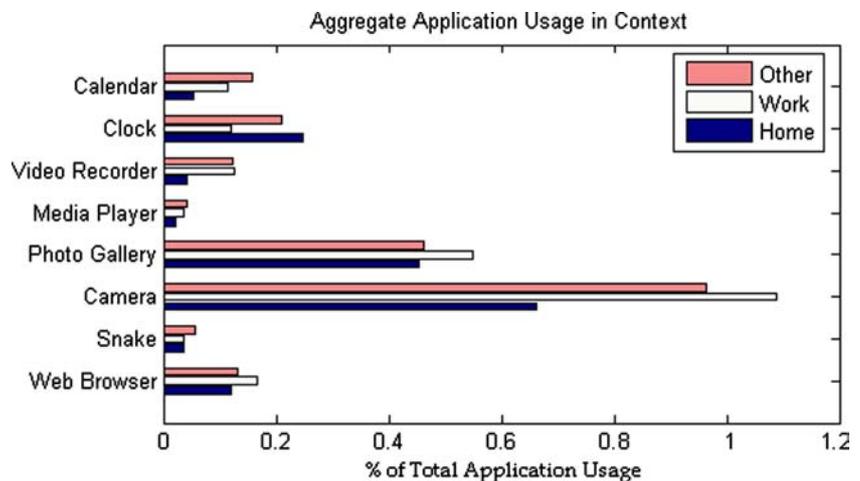
Fig. 9 Average communication mediums for 90 subjects (approximately 10 of the subjects did not use the phones as a communication device and were excluded from this analysis). The color bar on the right indicates the percentage each communication medium (Voice, Text, and Data) is used. All subjects use voice as the primary means of communication, while about 20% also actively use the data capabilities of the phone. Less than 10% of the subjects send a significant number of text messages

Fig. 10 Movement and communication visualization of the Reality Mining subjects. In collaboration with Stephen Guerin of Redfish Inc, we have built a Macromedia Shockwave visualization of the movement and communication behavior of our subjects. Location is based on approximate location of cell towers, while the links between subjects are indicative of phone communication

identified, destroying portions of the data collected during the months of September and October for six Sloan students and four Media Lab students.

3.4.2 Bluetooth errors

One central intent of this research is to verify the accuracy of automatically collected data from mobile phones for quantifying social networks. We are facing several technical issues. The 10-m range of Bluetooth, along with the fact that it can penetrate some types of walls, means that people who are not physically proximate may incorrectly be logged as such. By scanning only periodically every 5 min, shorter proximity events may also be missed.

Additionally, there is a small probability (between 1 and 3% depending on the phone) that a proximate, visible device will not be discovered during a scan. Typically this is due to either a low level Symbian crash of an application called the "BTServer," or a lapse in the device discovery protocol. The BT server crashes and restarts approximately once every three days (at a 5-min scanning interval) and accounts for a small fraction of the total error. However, to detect other subjects, we can leverage the redundancy implicit in the system. Because both of the subjects' phones are actually scanning, the probability of a simultaneous crash or device discovery error is less than 1 in 1,000 scans.

In our tests at MIT, we have empirically found that these errors have little effect on the correlations between interaction (survey data) and the 10 m Bluetooth proximity information. These problems therefore produce a small amount of 'background noise' against which the true proximity relationships can be reasonably measured. However, social interactions within an academic institution are not necessarily typical of a broader cross-section of society, and the errors may be more severe or more patterned. If testing in a more general population shows that the level of background noise is unacceptable, there are various technical remedies available. For instance, the temporal pattern of BTID logs allows us to identify various anomalous situations. If someone is not involved in a specific group conversation but just walking by, then they will often enter and leave the log at a different time than the members of the group. Similar geometric and temporal constraints can be used to identify other anomalous logs.

3.4.3 Human-induced errors

The two primary types of human-induced errors in this dataset result from the phone either being off, or separated from the user. The first error comes from the phone being either explicitly turned off by the user or exhausting the batteries. According to our collected survey data, users report exhausting the batteries approximately 2.5-times each month. One-fifth of our subjects manually turn the phone off on a regular basis during specific contexts such as classes, movies, and (most frequently) when sleeping. Immediately before the phone powers down, the event is timestamped and the most recent log is closed. A new log is created when the phone is restarted and again a timestamp is associated with the event.

A more critical source of error occurs when the phone is left on, but not carried by the user. From surveys, we have found that 30% of our subjects claim to never forget their phones, while 40% report forgetting it about once each month, and the remaining 30% state that they forget the phone approximately once each week. Identifying the times where the phone is on, but left at home or in the office presents a significant challenge when

working with the dataset. To grapple with the problem, we have created a 'forgotten phone' classifier. Features included staying in the same location for an extended period of time, charging, and remaining idle through missed phone calls, text messages, and alarms. When applied to a subsection of the dataset which had corresponding diary text labels, the classifier was able to identify the day where the phone was forgotten, but also mislabeled a day when the user stayed home sick. By ignoring both days, we risk throwing out data on out-lying days, but have greater certainty that the phone is actually with the user. A significantly harder problem is to determine whether the user has temporarily moved beyond 10 m of his or her office without taking the phone. Casual observation indicates that this appears to happen with many subjects on a regular basis and there are not enough unique features of the event to classify it accurately. However, as discussed in the relationship inference section, while frequency of proximity within the workplace can be useful, the most salient data come from detecting a proximity event outside MIT, where temporarily forgetting the phone is less likely to repeatedly occur.

3.4.4 Missing data

Because we know when each subject began the study, as well as the dates that have been logged, we can know exactly when we are missing data. This missing data is due to two main errors discussed above: data corruption and powered-off devices. On average we have logs accounting for approximately 85.3% of the time since the phones have been deployed. Less than 5% of this is due to data corruption, while the majority of the missing 14.7% is due to almost one-fifth of the subjects turning off their phones at night.

3.4.5 Surveys and diaries vs. phone data

In return for the use of the Nokia 6600 phones, students have been asked to fill out web-based surveys regarding their social activities and the people they interact with throughout the day. Comparison of the logs with survey data has given us insight into our dataset's ability to map accurately social network dynamics. Through surveys of approximately 40 senior students, we have validated that the reported frequency of (self-report) interaction is strongly correlated with the number of logged BTIDs ($R = 0.78$, $p = 0.003$), and that the dyadic self-report data has a similar correlation with the dyadic proximity data ($R = 0.74$, $p < 0.0001$). Interestingly, the surveys were not significantly correlated with the proximity logs of the incoming students. Additionally, a subset of subjects kept detailed activity diaries over several months. Comparisons revealed no systematic errors with respect to proximity and location, except for omissions due to the phone being turned off.

4 Community structure: complex social systems

In the previous section we showed that Bluetooth-enabled mobile phones might be used to discover a great deal about the user's patterns of activity. In this section we will extend this base of user modeling to explore modeling complex social systems.

By continually logging and time-stamping information about a user's activity, location, and proximity to other users, the large-scale dynamics of collective human behavior can be analyzed. If deployed within a group of people working closely together, correlations between the phone log and proximity log could also be used to provide insight behind the factors driving mobile phone use. Furthermore, a dataset providing the proximity patterns and relationships within large groups of people has implications within the computational epidemiology communities, and may help build more accurate models of airborne pathogen dissemination, as well as other more innocuous contagions, such as the flow of information.

4.1 Human landmarks

As shown in Figs. 4 and 12, there are people who users only see in a specific context (in this instance, at work). If we know the user is at work, information about the time of day, and optionally the location within the building (using static Bluetooth devices) can be used to calculate the probability of that user seeing a specific individual, by the straightforward application of Bayes' rule.

In contrast to previous work that requires access to calendar applications for automatic scheduling [21], we can generate inferences about whether a person will be seen within the hour, given the user's current context, with accuracies of up to 90% for "low entropy" subjects. These predictions can inform the user of the most likely time and place to find specific colleagues or friends. We believe that the ability to reliably instigate casual meetings would be of significant value in the workplace. We must also remember, however, that the ability to predict people's movements can be put to less savory uses. Careful consideration must be given to these possibilities before providing free access to such data.

4.2 Relationship inference

In Sect. 3 we discussed how information about location and proximity can be used to infer a user's context. In much the same way, knowledge of the shared context of two users can provide insight into the nature of their association. For example, being near someone at 3 pm by the coffee machines confers different meaning than being near the person at 11 pm at a local bar. However, even simple proximity patterns provide an indication of
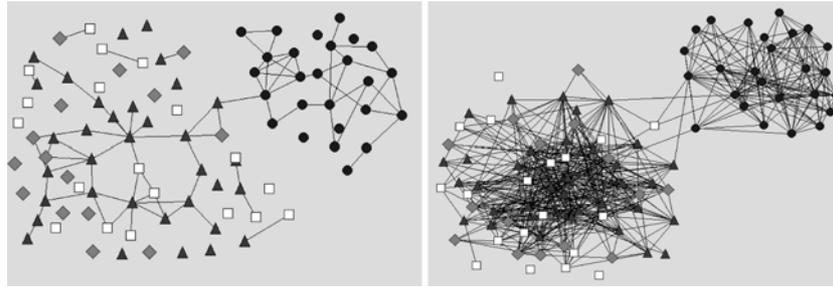
Fig. 11 Friendship (*left*) and daily proximity (*right*) networks. *Circles* represent incoming Sloan business school students. *Triangles*, *diamonds* and *squares* represent senior students, incoming students, and faculty/staff/freshman at the Media Lab. While the two networks share similar structure, inferring friendship from proximity requires the additional information about the context (location and time) of the proximity

the structure of the underlying friendship network as shown in Fig. 11. The clique on the top right of each network are the Sloan business students while the Media Lab senior students are at the center of the clique on the bottom left. The first year Media Lab students can be found on the periphery of both graphs.

We have trained a Gaussian mixture model [8] to detect patterns in proximity between users and correlate them with the type of relationship. The labels for this model came from a survey taken by all the experimental subjects at the end of two months of data collection. The survey asked who they spent time with, both in the workplace and out of the workplace, and who they would consider to be within their circle of friends. We compared these labels with estimated location (using cell tower distribution and static Bluetooth device distribution), proximity (measured from Bluetooth logs), and time of day.

Workplace colleagues, outside friends, and people within a user's circle of friends were identified with over 90% accuracy, calculated over the 2,000 potential dyads. Initial examination of the errors indicates that the inclusion of communication logs combined with a more

powerful modeling technique, such as Support Vector Machine, will have considerably greater accuracy.

Some of the information that permits inference of friendship is illustrated in Fig. 12 and Table 1. This figure shows that our sensing technique is picking up the common-sense phenomenon that office acquaintances are frequently seen in the workplace, but rarely outside the workplace. Conversely, friends are often seen outside of the workplace, even if they are co-workers. Determining membership in the 'circle of friends' requires cross-referencing between friends: is this person a member of a cluster in the out-of-office proximity data?

4.3 Proximity networks of work groups

By continuously logging the people proximate to an individual, we are able to quantify a variety of properties about the individual's work group. Although most work in networks assumes a static topology, proximity network data is extremely dynamic and sparse. We are currently building generative models to attempt to parameterize the underlying dynamics of these networks

Fig. 12 Proximity frequency data for a friend and a workplace acquaintance. The top two plots are the times (time of day and day of the week, respectively) when this particular subject encounters another subject he has labeled as a "friend." Similarly, the subsequent two plots show the same information for another individual the subject has labeled as "office acquaintance." It is clear that while the office acquaintance is encountered more often, the distribution is constrained to weekdays during typical working hours. In contrast, the subject encounters his friend during the workday, but also in the evening and on weekends
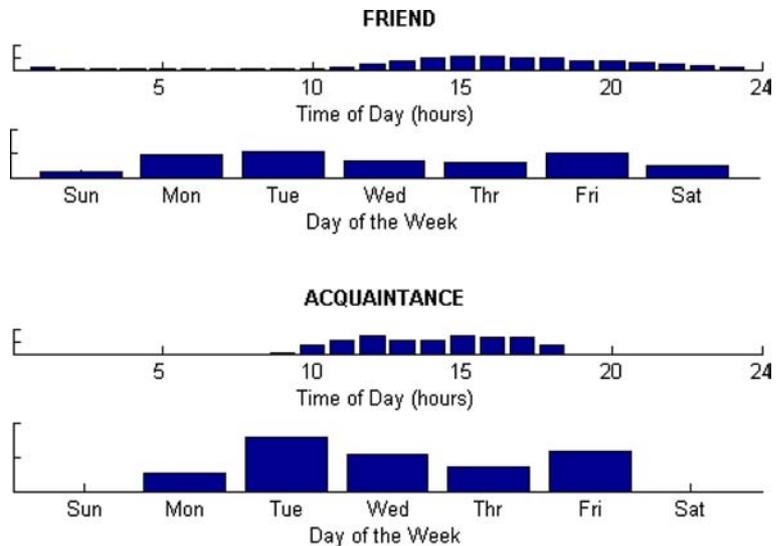
Table 1 Statistics correlated $(0.25 < R < 0.8, \quad p < 0.001)$ with friendship generated from 60 subjects (comprising 75 friendships) who work together at the Media Lab

	Friends		Not friends	
	Avg	Std	Avg	Std
Total proximity (min/day)	72	150	9.5	36
Saturday night proximity (min/week)	7.3	18	.20	1.7
Proximity with no signal (min/day)	12	20	2.9	20
Total number of towers together	20	36	3.5	4.4
Proximity at home (min/day)	3.7	8.4	.32	2.2
Phone calls/day	.11	.27	.001	.017

to gain insight into the functionality of the group itself. Additionally, we hope that by quantifying these proximity networks and contrasting the dynamics of the different groups at the Media Lab, we will gain some insight into the underlying characteristics of the research groups.

While each research group at the Media Lab is centralized around a faculty director, the proximity networks are not reflective of this static organizational structure. In many instances, the proximity network's degree distribution is indicative of a hub-and-spoke formation, however, the roles that are played within this structure are not static. Individuals that are hubs during one period of time fluidly exchange places with other team members on the periphery of the proximity network. This type of dynamic may be characteristic of the underlying nature of research groups at the Media Lab. As deadlines approach for specific individuals, they begin to spend more time in the Media Lab and increasingly rely on support from the rest of the group. Upon
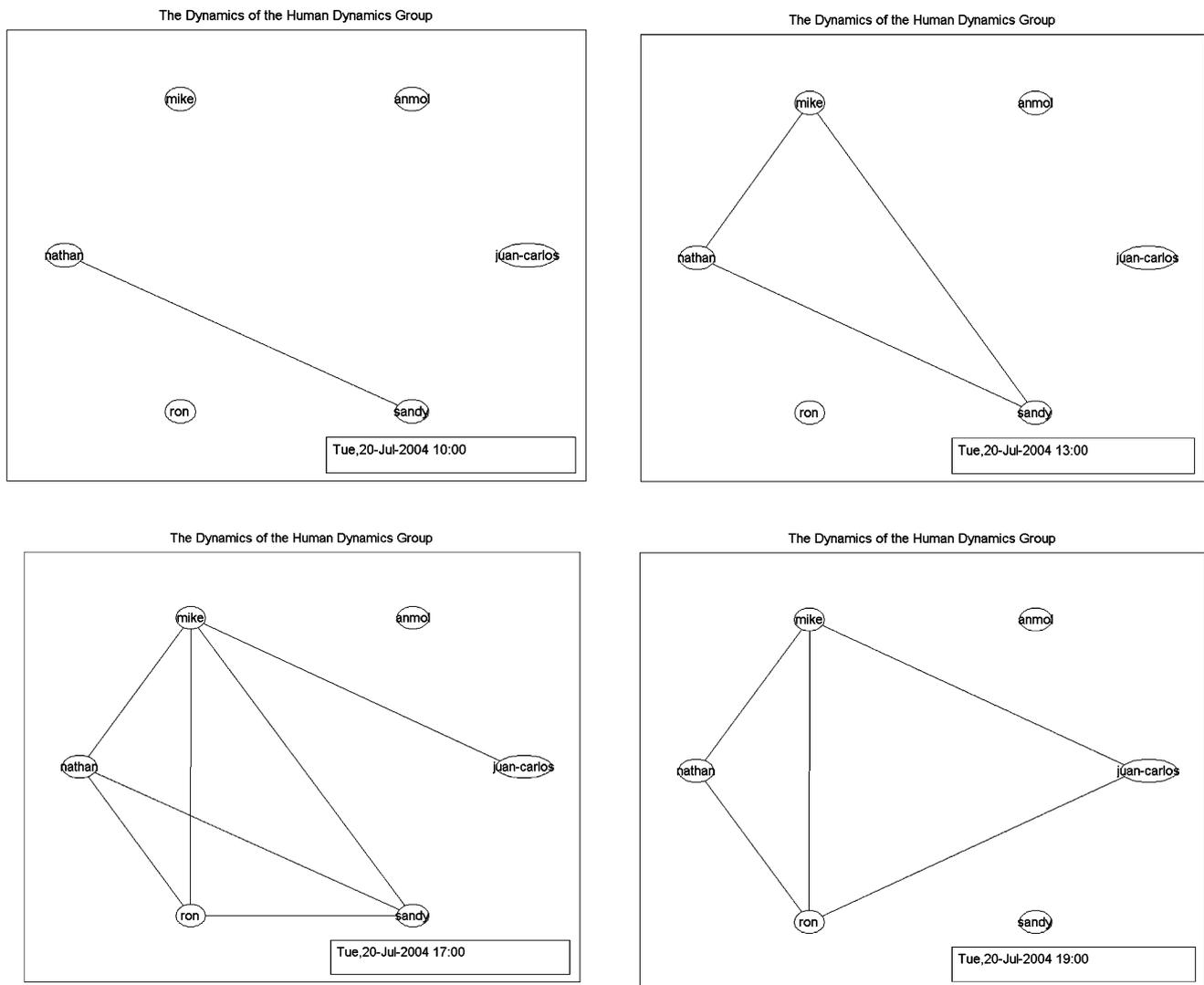

Fig. 13 Proximity network snapshots for a research group over the course of one day. In this example, if two of the group members are proximate during a 1-h window, an edge is drawn between them. The four plots represent four of these 1-h windows throughout the day at 10:00, 13:00, 17:00, and 19:00. We have the ability to generate these network snapshots at any granularity, with windows ranging from 5 min to three months
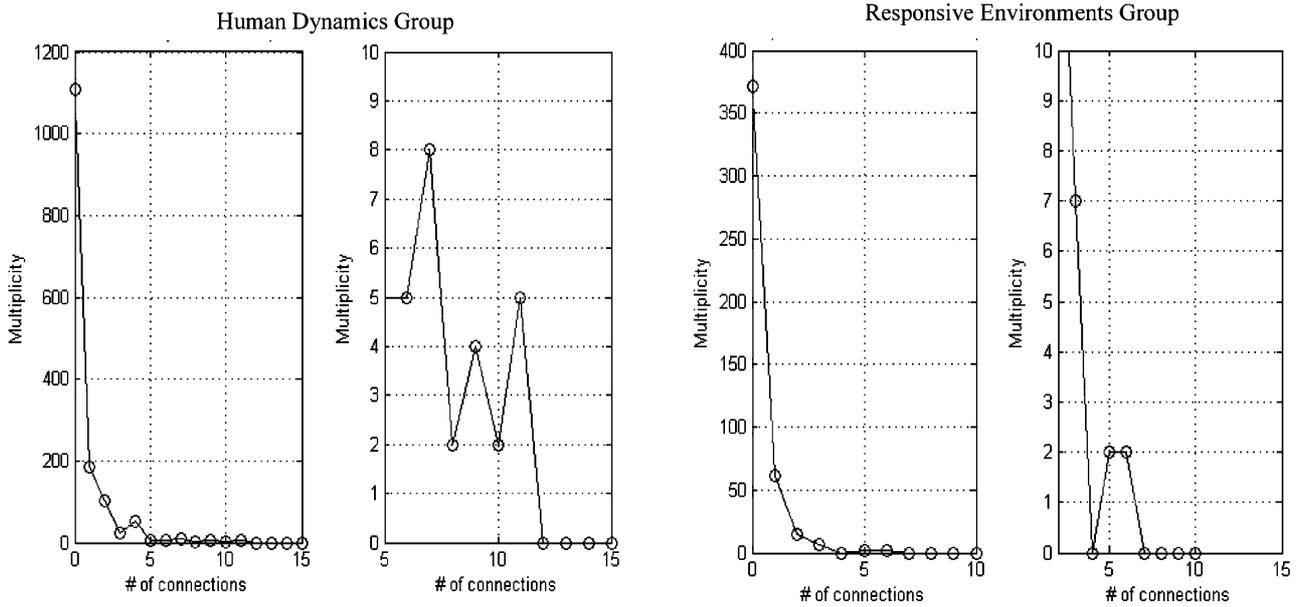
Fig. 14 Proximity Network Degree distributions between two groups. The *left-most plot* corresponds to the Human Dynamics group's degree distribution (i.e., the number of group members each person is proximate to over an aggregate of network snapshots). The *second left-most plot* is simply zoomed-in on the tail of the previous plot's distribution. Likewise, the *two right-most plots* are of the Responsive Environments group's degree distribution

Fig. 15 Proximity Time-Series and Organizational Rhythms. The *top plot* is total number of edges each hour in the Media Lab proximity network from August 2004 to January 2005. When a discrete Fourier transform is performed on this time series, the *bottom plot* confirms two most fundamental frequencies of the dynamic network to be (not surprisingly) 1 day and 7 days
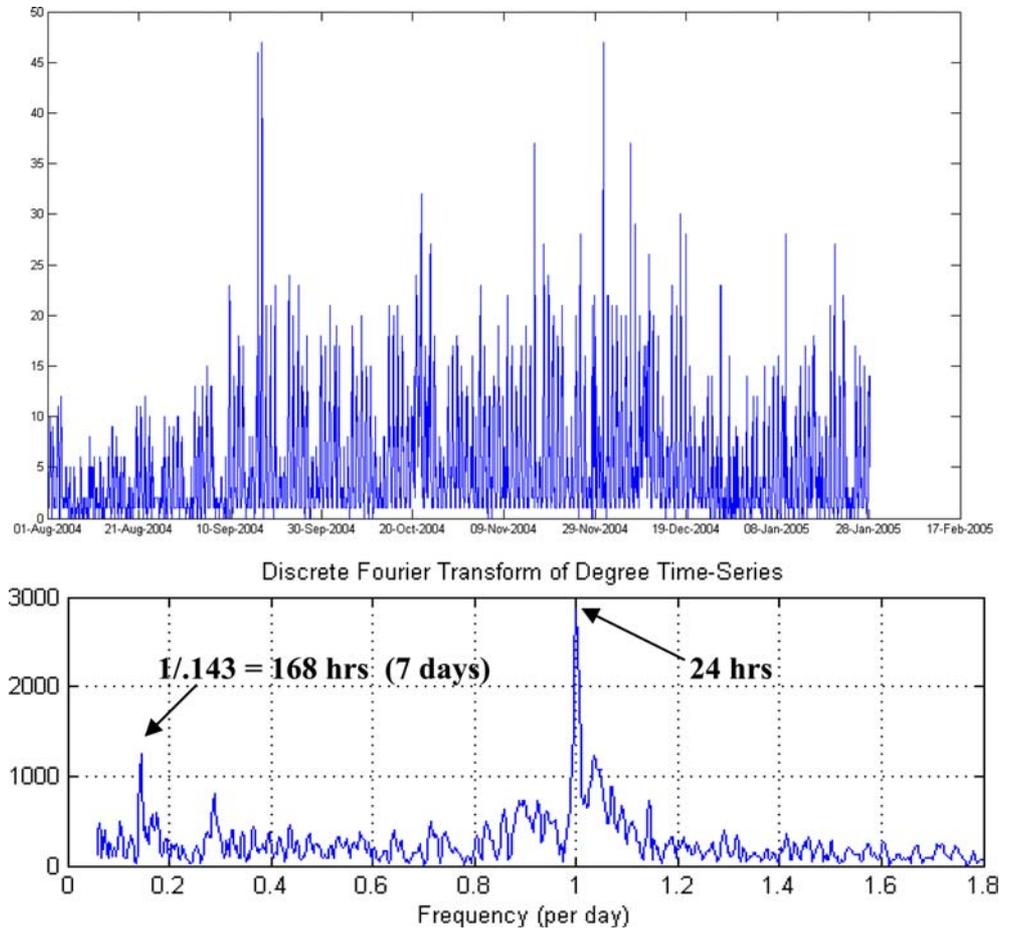
completion of a project, they resume their normal routines and can provide similar support to others. As will be discussed in the next section, this pattern of behavior has been shown to vanish when the entire group (or organization) is working towards the same deadline.

4.4 Organizational rhythms and network dynamics

Organizations have been considered microcosms of society, each with their own cultures and values. Similar to society, organizational behavior often shows recurrent patterns despite being the sum of the idiosyncratic behavior of individuals [5]. We are beginning to explore the dynamics of behavior in organizations in response to both external (stock market performance, a Red Sox World Series victory) and internal (deadlines, reorganization) stimuli (Figs. 13, 14).

During October, the 75 Media Lab subjects had been working towards the annual visit of the Laboratory's sponsors. Preparation for the upcoming events typically consumes most people's free time and schedules shift dramatically to meet deadlines and project goals. It has been observed that a significant fraction of the community tends to spend much of the night in the Lab finishing up last-minute details just before the event. We are beginning to uncover and model how the aggregate work cycles expand in reaction to these types of global deadlines. Figure 15 is a time series of the maximum number of links in the Media Lab proximity network during every 1-h window. It can be seen that the number of links in the Media Lab proximity network remained significantly greater than zero during the third week of October and in early December, representing preparation for a large Media Lab sponsor event and MIT's final week. It is possible to convert this time series into the frequency domain using a discrete Fourier Transform. The Fourier transform of this times series (Fig. 15, bottom) uncovers two fundamental frequencies, the strongest being at 24 h (1 day), and the second being at 168 h (7 days).

5 Conclusions

It is inevitable that the mobile devices of tomorrow will become both more powerful and more aware of their user and his or her context. We have distributed a fleet of one hundred context logging mobile phones throughout a laboratory and a business school at MIT. The data these devices have returned to us is unprecedented in both magnitude and depth. The applications we have presented include ethnographic studies of device usage, relationship inference, individual behavior modeling, and group behavior analysis. However, there is much more to be done, and it is our hope that this new type of data will inspire research in a variety of fields ranging from qualitative social science to theoretical artificial intelligence.

References

1. Addlesee M, Curwen R, Hodges S, Newman J, Steggles P, Ward A, Hopper A (2001) Implementing a sentient computing system. IEEE Comput Mag 34(8):50–56
2. Ashbrook D, Starner T, (2003) Using GPS to learn significant locations and predict movement across multiple users. Pers Ubiquit Comput 7:275–286
3. Bar-Noy A, Kessler I (1993) Tracking mobile users in wireless communication networks. IEEE Trans Inform Theory 39(6):1877–1886
4. Bhattacharya A, Das SK (1999) LeZi-update: an information-theoretic approach to track mobile users in PCS networks. In: Proceedings of the international conference on mobile computing and networking, Seattle, WA
5. Begole J, Tang J, Hill R (2003) Rhythm modeling, visualizations and applications. In: Proceedings of the UIST 2003. pp 11–20
6. Choudhury T (2004) Sensing and modeling human networks. Ph. D. Thesis, Program in Media Arts and Sciences, MIT, Cambridge, MA
7. Cox D, Kindratenko V, Pointer D (2003) IntelliBadge™: towards providing location-aware value-added services at academic conferences. In: Proceedings of UbiComp 2003, Lect Note Comput Sci 2864:264–280
8. Duda Ro, Hart PE, Stork DG (2001) Pattern classification. Wiley, New York
9. Eagle N, Pentland A (2005) Social serendipity: mobilizing social software. In: IEEE pervasive computing, special issue: the smart phone, pp 28–34
10. Himberg J, Korpiaho K, Mannila H, Tikanmäki J, Toivonen HTT (2001) Time series segmentation for context recognition in mobile devices. In: Proceedings of the IEEE international conference on data mining (ICDM 2001), San José, pp 203–210
11. Holmquist LE, Falk J, Wigström J (1999) Supporting group collaboration with inter-personal awareness devices. J Pers Technol 3(1–2)
12. Kim SJ, Lee CY (1996) Modeling and analysis of the dynamic location registration and paging in microcellular systems. IEEE Trans Vehicular Technol 45(1):82–90
13. Laasonen K, Raento M, Toivonen H (2004) Adaptive on-device location recognition. In: Proceedings for pervasive, pp 287–304
14. Laibowitz M (2004) Parasitic mobility for sensate media, MS Thesis in Media Arts and Sciences, MIT, Cambridge
15. LaMarca A, Chawathe Y, Consolvo S, Hightower J, Smith I, Scott J, Sohn T, Howard J, Hughes J, Potter F, Tabert J, Powledge P, Borriello G, Schilit B (2005) Place lab: device positioning using radio beacons in the wild. In: Proceedings of the third international conference on pervasive computing, May 2005
16. Lamming M, Newman W (1992) Activity-based information retrieval: technology in support of personal memory. IFIP Transact A-14:68–81
17. Liao L, Fox D, Kautz H (2004) Learning and inferring transportation routines. In: Proceedings for the national conference on artificial intelligence (AAAI-04), San Jose, CA
18. Mäntyjärvi J, Himberg J, Kangas P, Tuomela U, Huuskonen P (2004) Sensor signal data set for exploring context recognition of mobile devices. In: Workshop "Benchmarks and a database for context recognition" in conjunction with the 2nd international conference on pervasive computing (PERVASIVE 2004), April 18–23, Linz/Vienna, Austria

19. Paulos E, Goodman E (2004) The familiar stranger: anxiety, comfort, and play in public places. In: Proceedings of CHI 2004, pp 223–230
20. Raento M, Oulasvirta A, Petit R, Toivonen H (2005) ContextPhone—a prototyping platform for context-aware mobile applications. IEEE Pervasive Comput 4(2):51–59
21. Roth J, Unger C (2000) Using handheld devices in synchronous collaborative scenarios. In: Proceedings of the second international symposium on handheld and ubiquitous computing (HUC), Bristol, UK
22. Schilit B, Adams N, Gold R, Tso M, Want R (1993) The ParcTab mobile computing system. In: Proceedings of the fourth workshop on workstation operating systems, pp 34–39
23. Singh P, Williams W (2003) LifeNet: a propositional model of ordinary human activity. In: Proceedings of the workshop on distributed and collaborative knowledge capture (DC-KCAP) at K-CAP 2003
24. Want R, Hopper A, Falcao V, Gibbons J (1992) The active badge location system. ACM Trans Inform Syst 10:91–102
25. Wolf J, Guensler R, Bachman W (2001) Elimination of the travel diary: an experiment to derive trip purpose from GPS travel data. In: Proceedings from the transportation research board 80th annual meeting, Washington DC
26. Weld D, Anderson C, Domingos P, Etzioni O, Gajos K, Lau T, Wolfman S (2003) Automatically personalizing user interfaces. In: Proceedings of the international joint conference on artificial intelligence (IJCAI03), Acapulco, Mexico
27. http://news.bbc.co.uk/1/hi/business/4257739.stm (verified April 13, 2004
28. http://www.bluetoothsig.com/emailfiles/Zelos.pdf (verified April 13, 2004)