Reality Mining Africa

Shawndra Hill[1], Anita Banser[2], Getachew Berhan[3], Nathan Eagle[4]
Operations and Information Management Department[1,2], Department of Information Science[3], Santa Fe Institute[4]
University of Pennsylvania, Philadelphia, PA[1,2], Addis Ababa University, Ethiopia[3], Santa Fe Institute, Santa Fe, New Mexico[4]
shawndra@wharton.upenn.edu [1], bansera@seas.upenn.edu [2], getachewb1@yahoo.com [3], nathan@mit.edu [4]

Abstract

Cellular phones can be used as mobile sensors, continuously logging users' behavior including movement, communication and proximity to others. While it is well understood that data generated from mobile phones includes a record of phone calls, there are also more sophisticated data types, such as Bluetooth or cell tower proximity logging, which reveal movement patterns and day-to-day human interactions. We explore the possibility of using mobile phone data to compare movement and communication patterns across cultures. The goal of this proof-of-concept study is to quantify behavior in order to compare different populations. We compare our ability to predict future calling behavior and movement patterns from the cellular phone data of subjects in two distinct groups: a set of university students at MIT in the United States and the University of Nairobi in Kenya. In addition, we show how Bluetooth data may be used to estimate the diffusion of an airborne pathogen outbreak in the different populations.

Introduction

New forms of communication, such as email, mobile phones, and instant messaging, have revolutionized the human behavioral data available to social scientists and computer scientists alike. In particular, the global explosion of mobile phone usage has opened an avenue to exploring population-level movement dynamics (Gonzalez, Hidalgo, and Barabasi 2008). Previously, surveys and other tools were required to assess the movement patterns of individuals within human communities. The utility of such instruments was threatened by responder and interviewer biases, as well as by data sparseness and discontinuity. In contrast, cellular phones can serve as continuous mobile sensors to continuously collect location and proximity information, such as by logging local cellular towers or Bluetooth devices within the scanning range of the user (Eagle and Pentland 2006; Eagle 2008). The use of specially equipped mobile phones eliminates many of the difficulties associated with less sophisticated data collection tools.

Mobile phones can also serve a communication log, by generating records of phone calls, short messaging, and packet data transfer information. This information can be compiled to provide insight into interaction patterns, which can be of particular interest, for instance, when comparing developed and non-developed nations. Communication data mining methods might also permit researchers to quantitatively compare human social responses to outlier events (e.g., natural disasters and political unrest).

Here we analyzed mobile phone data from subjects in the United States and Kenya, with the goal of comparing our ability to predict movement and future calling behavior in the two subject pools. We first determined descriptive statistics from the data, including information about the network structure. We then compared our ability to apply location (i.e., cell tower) or link (i.e., phone call) prediction methods in both datasets. Finally, we derived an interaction network from Bluetooth proximity data to compare the percolation methods of disease outbreaks in both populations. One can imagine using proximity data to estimate the amount of an individual's exposure to a disease given their social network and human contact on it. To our knowledge, this is the first work to compare the social network behavior and movement patterns of Americans and Africans at an individual level. Based on the results of this proof-of-concept analysis, we believe that the simple approach could be used to compare the movement and communication patterns of myriad cultures. The lack of sophisticated statistical relational learning techniques to incorporate social network, location, and temporal data at a large scale, however, provides an opportunity for machine learning researchers to develop dynamic graphical approaches for link and location prediction.

Reality Mining Data

Data collection software, as described in (Raento et al. 2005), was installed on the mobile phones provided to students at universities in both the United States and Kenya. Kenyan call data was collected from 38 [1] users from October 2006 to May 2008, and Kenyan location data was collected from 38 users from January 2006 to May 2006 and from October 2006 to May 2008. US call data was collected from 90 users from July 2004 to January 2005, and US location data was collected from 77 users from January 2004 and July 2004 to Jan 2005. Kenyan participants were

[1] 72 users signed up for the study but only 38 uploaded data

second-year computer science students at the University of Nairobi. US subjects were students at MIT. All data were anonymized to ensure the privacy of participants.

Call log details collected from the phones included the following data for each communication: the type of communication (voice call, text message, packet data, etc.), the unique user ID of the cell phone user and that of the person with whom the user communicated, and the timestamp (start, end, and duration times) of the communication. Cell tower proximity log details included the unique user ID of the cell phone user, the timestamp at which the data was collected, and the ID of the cell towers within the cell phone's detectable range. Likewise, Bluetooth proximity log details included the Bluetooth IDs of users who were close at a given timestamp. These three types of data, call details, cell tower locations, and Bluetooth IDs, were used to construct networks of interactions.

Methods

Predicting Call Links

People can be identified by the people they call (Hill et al. 2006). However, the extent to which this is true relies on limited change to an individual's calling circle. To quantify this change, we calculated the amount of overlap in friends' links from month to month for US and Kenyan users using the Dice similarity score (Dice 1945):

$$D(A, B) = \frac{2|A \cap B|}{|A| + |B|}, \tag{1}$$

where A is the set of links in time period one (the training period), and B is the set of links in time period two (the test period). In this way, we observed how well the links in the training period correspond to those in the test period. The Dice criterion is bounded between 0 and 1, with a value of 1 when the sets are identical and 0 when there is no overlap. In addition to the Dice criterion, we used the absolute score determined as the number of call link overlaps from month to month. For example, for November-December 2004, the overlap score per user is the number of people called by a user in both November and December. We applied this technique across multiple months to avoid sensitivity to the test period.

Predicting Location

By logging timestamped cell tower locations, we were able to construct individual movement models. Our goal was to predict whether the user was at home, work, or somewhere else at a given time. We chose to examine data from the month of October (Kenya: October 2007; US: October 2004) because it falls during the school semester (and thus students are expected to be at school) and because October does not have any holidays, which might cause deviation from 'normal' cell phone user behavior. We inferred user location (i.e., home, work, other, or no signal) based on the visible tower(s):. 'work' towers were identified that were registered by $> 20\%$ of users' cell phones during October 2004 (US) or October 2007 (Kenya) during customary working hours (0900-1700, Monday to Friday). This method was chosen because most subjects are students therefore likely work at the same location. We then identified the top 8 towers that showed up for individual users during the work hours and labeled them Work. In addition to labeling all $group$ level work towers as Work, we also labeled the $individual$ level user-tower ID pairs that showed up in the top 8 towers from above as Work. Individual 'Home' towers were identified as the top 8 towers that showed up for individuals from 0200-0400 (Sunday to Sunday), based on the assumption that most users were at home during these hours. User-tower ID pairs that were not already labeled Work were also labeled Home. All other unlabeled user-tower ID pairs were labeled 'Other'. To label each hour of a given day, we used a maximum vote method: That is, for every hour, we labeled the location by taking the most frequent label in that hour. We labeled hours for which there were no registered tower location data as 'No Signal' hours. The towers or locations learned in one time period were used to predict locations in another. We built a decision tree model to predict user location conditioned on the hour of day and day of the week. In addition to the decision tree model, we used Shannon entropy to quantify and compare movement behavior:

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_b p(x_i), \tag{2}$$

We approached the entropy calculation as a four class problem, where b is set to 4, and $p(x_i)$ is the probability of person i being at Home, Work, or Other or having No Signal.

Predicting Potential Disease Transmission

The Bluetooth data enabled us to construct interaction networks of subjects. Two users form a link when they are in close proximity to one another. These data may be useful in predicting the spread of a virus. There has been recent work linking wireless network data to virus propagation (Khayam and Radha 2005; Zheng, Li, and Gao 2006). However, the focus is typically on software viruses and not communicable diseases. We apply a well-discussed epidemiology theory of disease transmission, the Susceptible-Immune-Recovery model, where subjects are assumed a level of susceptibility, immunity, and ability to recover (Hethcote 1976). We used the publicly available software sispread (Alvarez et al. 2010) to simulate outbreaks on our Bluetooth network.

Results

Descriptive Statistics

Location and call data from US and Kenyan users were compared using visualization and descriptive statistics of their social network including degree distribution and clustering coefficient. We plotted trends by the time of day for the different communication types (i.e., voice call, short message, or packet data) for both Kenyan and US data.

In terms of communication method, US users sent markedly more packet data messages than Kenyan users, and Kenyan users made more voice calls than US users. There was no significant difference in short messaging use (Fig. 1).

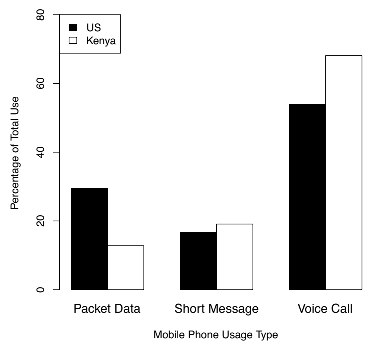

Figure 1: Communication types for US and Kenyan mobile phone users

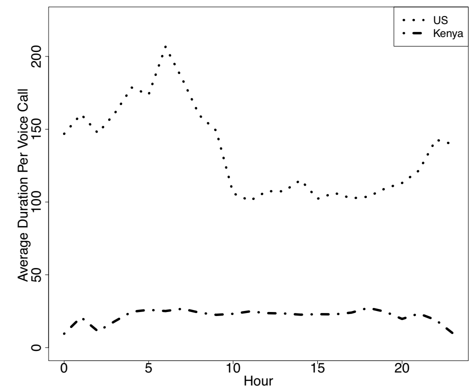

Figure 3: Average mobile phone call duration according to the hour of day for US and Kenyan users

These results may indicate that fewer Kenyan users utilized or had access to packet data mobile phone plans, which is consistent with Kenya's being a less-developed nation than the United States. The Kenyan data displayed a marked difference between the average number of voice calls and average number of unique voice calls (data not shown), and US users consistently made fewer and longer voice calls than Kenyan users (Figs. 2 and 3, $P < 0.0001$). While US users exhibited a peak in call duration in the mornings at approximatly 0800, the call duration for Kenyan users remained relatively constant (Fig. 3). The duration of packet data calls (downloads) was also significantly higher in the United States.

These data may suggest that dropped calls and/or poor signaling were more common among Kenyan users, potentially suggesting a poorer quality of service overall or a sparser cell tower distribution (although the tower distribution in Nairobi is reportedly similar to the distribution in Boston). Alternatively, the increased number of "missed calls" may indicate a difference in economic development between the two user populations (Fig. 4). Mobile phone users residing in low-income areas frequently use "missed calls" to relay a message. For example, a cell phone user may purposefully place a "missed call" to indicate that he or she has returned home safely, thus relaying a predetermined message without using airtime. Outbound calls from Kenyan users were either missed or had zero duration 62.01% of the time compared to 31.8% US .

While most US voice calls were made in the evenings (after 1700), Kenyan users had the highest mobile phone usage during work hours (Fig. 5), suggesting differences between the work cultures of the two countries or that different types of calls (work-related vs. personal calls) were made at different times of day. The call duration also varied depending on the day of the week in both the United States and Kenya. For US users, the call duration peaked on Wednesday and Thursday at 1000 and 2300. For Kenyan users, call duration peaked on Tuesday at 1400 (data not shown).

Figure 6 shows the distribution of unique towers seen by Kenyan and US users as a function of time, which can serve as a proxy for general movement. Together with the other data, this figure suggests that while the movement patterns of US and Kenyan users were similar, their overall call patterns were not.

The Kenyan and US user within-subject networks were relatively similar with respect to degree distribution (Fig. 7). This result indicated that each Kenyan/US user called the same number of people as other Kenyan/US users (on

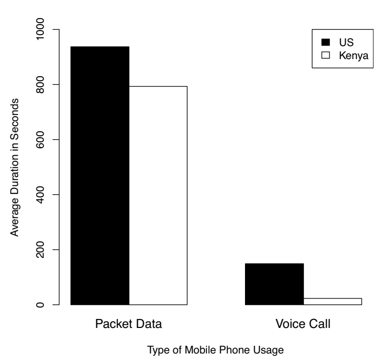

Figure 2: Mobile phone call/download durations for Kenyan and US users.

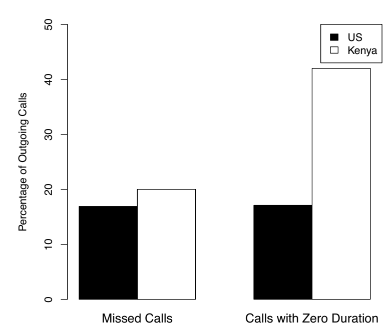

Figure 4: Percentage of all outgoing calls that are either missed or have zero call duration

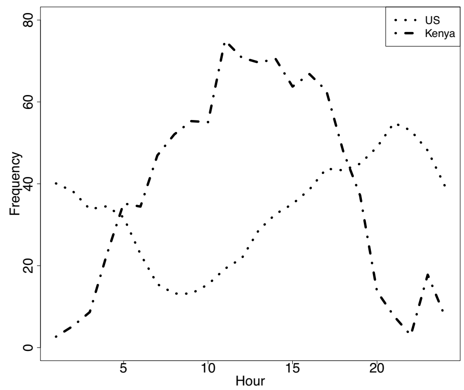

Figure 5: Distribution of total voice calls per person according to hour of day. Data are for July 04-Nov 04 (US) or July 07-Nov 07 (Kenya). Similar distributions were observed for short messages.

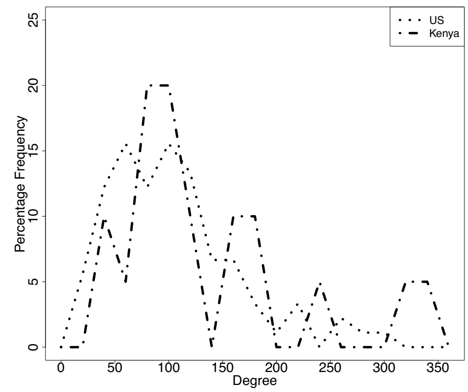

Figure 7: Degree distributions for voice calls for US and Kenyan users. Data are for Jul 04 - Jan 05 (US) and Jul 07 - Jan 08 (Kenyan).

average). Furthermore, phone call network plots (Fig. 8) indicated that users within both the US and Kenya were highly clustered with their colleagues.

Predicting Call Links

Because the call data was very sparse, to compare call behavior across two months, we isolated the months for which Kenyan and US data had $> 75\%$ of users making calls in two consecutive months. These 'overlap months' were November and December 2004 (United States) and November and December 2006 (Kenya). However, we applied our approach to all possible month pairs in our dataset.

In Figure 9, we plotted data using both the Dice coefficient score and absolute overlap score (number of friends in common). Dice score distribution plots indicate the self-similarity of users called when each person in one month is compared to themselves in the next month. A higher mean suggests that people are more "stable" in whom they call, while a lower mean suggests less "stability" (i.e., more change in behavior from month to month). Based on all calls and calls for the top 10 friends (ranked by number of minutes communicated with during the first month), the Kenyan group was less stable by Dice score (Fig. 9).

Predicting Location

We next assessed how well we could predict 'place' of a subject at any given time in our dataset. The attributes in our model were the day of the week and hour of the day, and the target was the place (Home, Work or Other) where the user was found. Our predictive models were evaluated based on their ability to rank individuals according to different attribute subsets. We used the area under the ROC curve (AUC) to compare models.

The average AUC scores for Home, Work, and Other for US and Kenyan users are shown in Table 1. It was relatively easier to predict Home than Work or Other locations. We also observed a large relative difference in AUC in our ability to predict Home compared to Work for Kenyan subjects - suggesting home life is much more stable (users arrive home at a certain hour and stay there).

Finally, we calculated the average entropy per hour for the subjects in the two countries. The tower locations (Home, Work, Other, or No Signal) recorded by US or Kenyan users in October 2004 or October 2007, respectively, were monitored over the course of 7 consecutive days (168 h of loca-

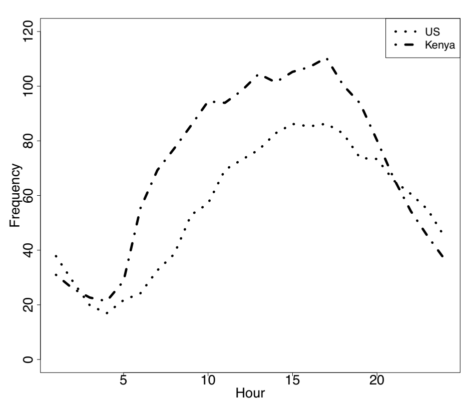

Figure 6: Distribution of the total number of towers seen by US and Kenyan users using the entire dataset.

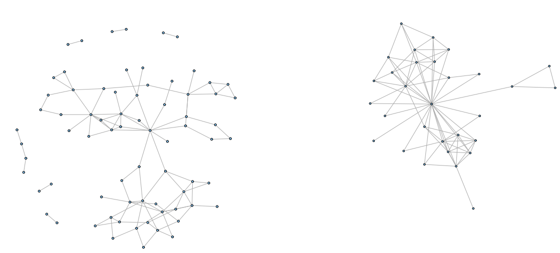

Figure 8: Within subject phone call links among Kenyan (left) and US (right) mobile phone users. Kenyan: 30 nodes, cc = 0.40; US: 73 nodes, cc = 0.23

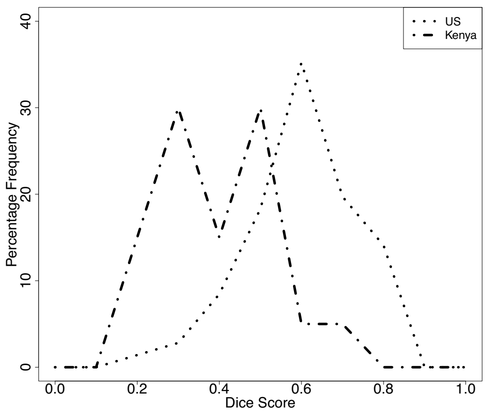

Figure 9: Dice coefficient plot showing the overlap score distribution for US and Kenyan users. Links between friends were established using both voice calls and short messages and then ranked by the number of minutes on the links to get the top 10 friends for comparison.

	US	Kenya
Home	0.598	0.684
Work	0.579	0.629
Other	0.538	0.532

Table 1: Average AUC for Home, Work, and Other Cell towers for US and Kenyan subjects

tion data). This process revealed entropies for Kenyan and US users of 1.32 and 1.519, respectively. These data indicate increased movement among US users, as compared to Kenyan users. It should be noted that the sample size for Kenyan users was very small (9 active Kenyan vs. 55 active US subjects).

Predicting Potential Disease Transmission

We applied SI, SIS, and SIR models (data not shown) to the Bluetooth networks - one for US data and one for Kenyan data - for 100 hypothetical time points. We found 52 subjects in the Kenyan data and 90 in the US. We sampled the US data to obtain 52 nodes (subjects) for comparison to the Kenyan data. The Kenyan and US users had clustering coefficients of 0.710 and 0.738, respectively, indicating that over our sample, almost all users came in contact (proximity) with one another at some point. Using sispread (Alvarez et al. 2010), we were able to run a number of simulations. In all simulations, the US data had a much higher average prevalence of disease with time than the Kenyan data. Because the sispread software does not use the actual network interaction frequency, instead assuming a uniform interaction on all links, we developed a tool to apply an SIR version that weights the internodal links. The weight between any two people is simply the number of times they come in contact (Bluetooth proximity) in the sample period. Based on our assumptions, we again found that the infection would spread quicker in the US network. Fig. 10 shows the estimated number of susceptible, infected, and recovered individuals at each time point. The Kenyan Bluetooth network was constructed with information gathered over 561 days, while the US Bluetooth network was constructed with information gathered over 300 days. Because of this, the number of in-network interactions, represented as edge weights, in the Kenyan Bluetooth network was disproportionately higher than that in the US network. To directly compare the networks, we normalized (multiplied) the Kenyan edge weights by the ratio 300/561. Before the workshop presentation, we intend to apply a weighted SIR network that also accounts for timestamps, rather than assuming everyone has an interaction at each time point.

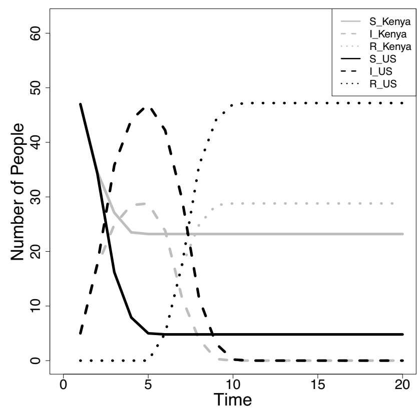

Figure 10: Number of estimated susceptible (S), infected (I) and recovered (R) people when incorporating the weights/number of Bluetooth interactions

Changes with Outlier Events

The call data included communications made during Christmas (December 25) and before, during, and after the 2007 Kenyan presidential elections that resulted in political unrest (December 26, 2007). Both of these events served as outlier events or exogenous ”shocks” that could have altered user communication habits or movements. Therefore, these time periods were of particular interest in assessing our ability to detect alterations in the communication dynamics.

The average mobile phone call duration among Kenyan users was longer during the Kenyan presidential elections (December 2007). During the election period, Kenyan users displayed increased voice calls, unchanged short message transfer, and reduced packet data transfer, as compared to before and after the election period (Fig. 11).

There were also strong shifts in the US data around Christmas and the New Year in voice calls vs. packet data, indicating that users sent more data and made fewer calls during that time (Fig.12). These results perhaps suggest that voice calls were generally made to call family members and that the students were at home (rather than at school) during the holidays. There seemed to be a similar shift among Kenyan users (albeit less pronounced), with an increased number of text messages vs. voice calls.

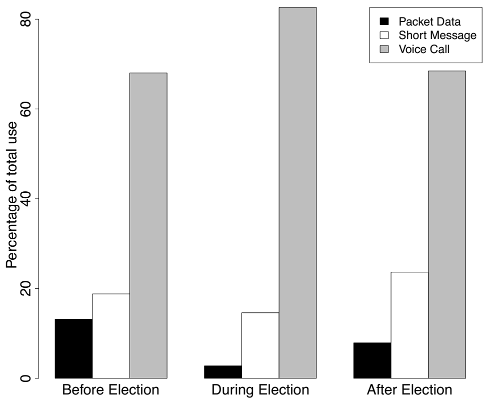

Figure 11: Percentage of communication use, before, during, and after the presidential election period

Discussion

There were a number of limitations in our study. Due to data collection issues, many of the Kenyan subjects only participated in one or two data collection sessions – resulting in shorter total duration and less subjects overall. This made it more challenging to draw specific conclusions in our comparisons between Kenyan and US mobile phone users, especially around predictably - which requires continuous, longitudinal data. These data constraints made it difficult to make comparisons about events of interest. For example, only 5 users were active in the Kenyan data during the entire 3-month period surrounding the election. Additionally, because of the election timing, the impact of the holiday season vs. elections is difficult to disambiguate. Despite these limitations, this paper represents the first comparison of mobile phone behavioral data between subjects from both Kenya and the US. We have demonstrated how mobile phone data from these very different regions can be used to gain better insight into the underlying dynamics of the studied population, ranging from typical movement and communication behavior to behavior during outlier events. With a larger sample population and longer experimental time, in-depth studies could be completed on the social network structures of different communities, and how these communicates evolve over the academic year. In future work, we will be continuing to collect Reality Mining data in effort to improve our understanding of people's similarities and differences across cultures and continents. In addition, future steps should include the development of machine learning methods to accommodate multidimensional network, spatial, and temporal data.

References

Alvarez, F. P.; Crepey, P.; M., B.; and Valleron, A. J. 2010. sispread: A software to simulate infectious diseases spreading on contact networks. *Methods of Information in Medicine* 1:19–26.

Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology* 26(3):297–302.

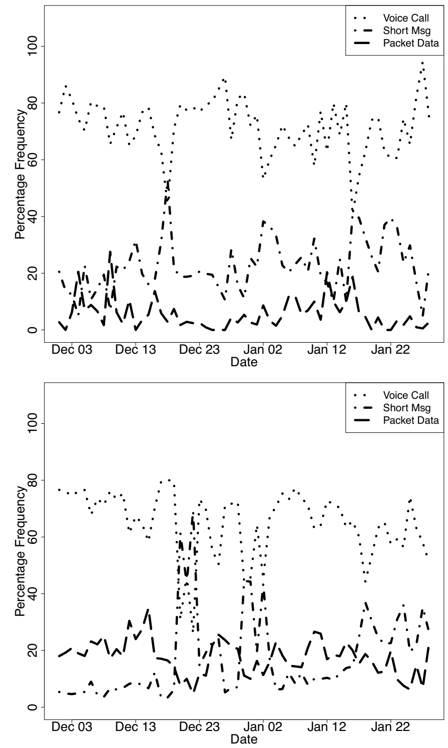

Figure 12: Percentage of all communications that were voice calls, short messages, or packet data as a function of date for Kenyan (top) and US (bottom) users

Eagle, N., and Pentland, A. 2006. Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing* 10(4):255–268.

Eagle, N. 2008. Behavioral inference across cultures: Using telephones as a cultural lens. *IEEE Intelligent Systems* 23(4):62–64.

Gonzalez, M.; Hidalgo, C.; and Barabasi, L. A. 2008. Understanding individual human mobility patterns. *Nature* 453:779–782.

Hethcote, H. 1976. Qualitative analyses of communicable disease models. *Math. Biosci.* 335–356.

Hill, S.; Agarwal, D.; Bell, R.; and Volinsky, C. 2006. Building an effective representation for dynamic networks. *Journal of Computational and Graphical Statistics* 15(3):584–608.

Khayam, S. A., and Radha, H. 2005. A topologically-aware worm propagation model for wireless sensor networks. *Proceedings of the 2nd Intl Workshop on Security in Distributed Computing Systems* 210–216.

Raento, M.; Oulasvirta, A.; Petit, R.; and Toivonen, H. 2005. Contextphone: A prototyping platform for context-aware mobile applications. *IEEE Pervasive Computing* 4:51–59.

Zheng, H.; Li, D.; and Gao, Z. 2006. An epidemic model of mobile phone virus. *Proceedings of the 1st Intl Symposium on Pervasive Computing and Applications* 1–5.