PREDICTABILITY OF MOBILE PHONE ASSOCIATIONS

Bjørn Sand Jensen, Jan Larsen, Kristian Jensen,
Jakob Eg Larsen, and Lars Kai Hansen

Technical University of Denmark
{bjje,jl,krije,jel,lkh}@imm.dtu.dk

Abstract. Prediction and understanding of human behavior is of high importance in many modern applications and research areas ranging from context-aware services, wireless resource allocation to social sciences. In this study we collect a novel dataset using standard mobile phones and analyze how the predictability of mobile sensors, acting as proxies for humans, change with time scale and sensor type such as GSM and WLAN. Applying recent information theoretic methods, it is demonstrated that an upper bound on predictability is relatively high for all sensors given the complete history (typically above 90%). The relation between time scale and the predictability bound is examined for GSM and WLAN sensors, and both are found to have predictable and non-trivial behavior even on quite short time scales. The analysis provides valuable insight into aspects such as time scale and spatial quantization, state representation, and general behavior. This is of vital interest in the development of context-aware services which rely on forecasting based on mobile phone sensors.

1 Introduction

The wide acceptance of sensor rich mobile phones and related applications enables deep studies of human behavior. According to a recent study by Song et al. [11, 12] based on mobile phone location trajectories, individual human mobility patterns are highly predictable. When including the complete history of the participants they derived an upper bound on prediction of the next location of 93% in a large cohort of 45,000 users. The upper bound is based on information theory. Using Fanos inequality it was shown in [12] that the entropy rate transforms into an upper bound of the predictability.

Interest in understanding human behavior using mobile technology is increasing, see e.g., the recent review by Kwok [9]. The work of Eagle et al. [4] on the Reality Mining dataset, marks an early and important contribution. Using a Hidden Markov model and the time of day, they demonstrate explicit prediction accuracies (home, work, elsewhere) in the order of 95%. Furthermore, they use principle component analysis (PCA) to visualize temporal patterns in daily life. The stability of these temporal patterns was confirmed by Farrahi et al. [5] in the same dataset using unsupervised topic models.

While Eagle et al. focus on finding statistical regularities in behaviors at the group level using parametric models Song et al. [11] are interested in individual predictability using non-parametric methods and argue that inter-participant variability is significant and in fact power-law distributed. For a further discussion on parametric and non-parametric models, and the relation to information theory, we refer to Bialek et al. [3].

We follow the implicit modeling approach by Song et al., i.e., using bounds rather than explicit predictors to discuss human behavior in a novel mobile phone data set that complements the analysis of Song et al. Our data set has significantly higher temporal resolution and involves more sensors, however, in a much smaller cohort ($N = 14$).

The opportunity to analyze multiple sensors is quite unique and produces new insight both on the predictability of each sensor and on sensor dependencies. We show that all the location and proximity based sensors have a relatively high predictability bounds for the entire population. Whereas Song et al. [11] provide important results on location prediction, the multiple sensors applied in our study can potentially provide a richer description of context beyond location. And as the extended set of sensors enjoys similar high predictability rates, it may contribute additional useful information on human behavior and support more general context dependent services.

Furthermore, we are interested in the predictability on different time scales, and to probe whether it is possible to predict non-trivial behaviors on smaller scales than the one hour time scale considered in [11]. Finally, we suggest applying mutual information - or prediction information [3] - as a method to easily estimate an optimal time scale in a non-parametric fashion. The information theoretic approach is based on upper and lower bounds on predictability. In this paper we use the upper bound proposed by Song et al., and derive and analyze a tighter lower bound based on a first Markov model [7], rather than the zero order model of [12].

An early version of this work was presented to the machine learning community in [7]. The present paper extends this work with; 1) an extended description of the experimental setup; 2) comparison between WLAN and GSM, in particular in relation to optimal time scale; 3) extended discussion and interpretation.

The paper first gives a presentation of the experimental setup and the acquired mobile phone dataset. In Section 3 we present the information theoretic tools necessary to follow the analysis of the dataset. The result of the study is presented and discussed in Section 4, followed by the conclusion in Section 5.

2 Experimental Setup

Within the last decade there has been a number of studies of real-world dataset or *lifelogs* reflecting human life [1, 2]. In this section we present a software platform for obtaining such *lifelogs* using standard off-the-shelf mobile phones functioning as individual wearable sensor platforms.

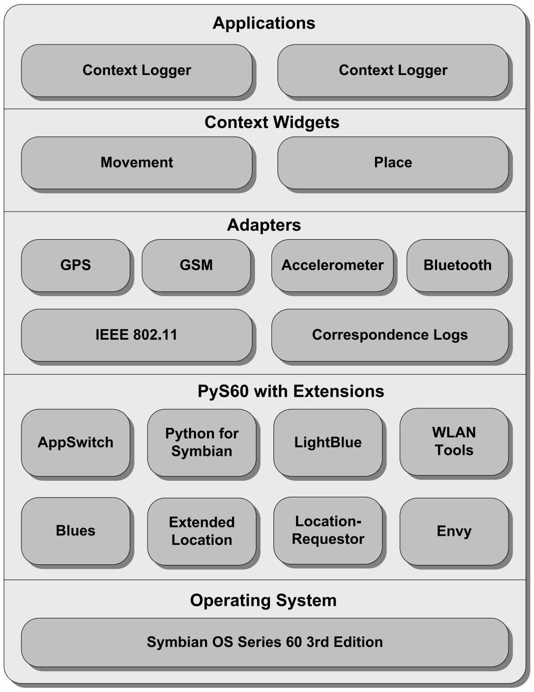


Fig. 1. Mobile Context Toolbox architecture [10]. The bottom two layers provide low-level access to the embedded sensors, whereas the adapters and context widget layers provide high-level Python interfaces to sensors and inferred information for applications.

2.1 Mobile Context Toolbox

Since utilizing multiple sensor inputs on mobile devices can be a complex task we have used our Mobile Context Toolbox [10], which provides an open extensible framework for the Nokia S60 platform running the Symbian mobile Operating System present in mobiles phones such as Nokia N95. The framework provides access to multiple embedded mobile phone sensors, including accelerometer, microphone, camera, etc., as well as networking components such as phone application data (calendar, address book, phone log, etc.), and phone state (profile, charge level, etc.).

In principle, mobile devices can acquire information from the surrounding environment as well as from online sources, but in the present study we focus on information that can be acquired through the large variety of sensors embedded in the device. The framework has multiple layers (as depicted in Fig. 1) on top of the Nokia S60 platform. The framework uses Python for S60 (PyS60) with a set of extensions for accessing low-level sensors and application data. The adapters layer provides interfaces for the low-level sensors, whereas the context widgets uses one or more adapters to infer higher-level contextual information. Finally, the application layer utilize contextual information inferred from context widgets and/or directly from context adapters. Further details and explanation of the Mobile Context Toolbox is provided in [10].

2.2 Logging Data From Embedded Sensors

For the purpose of using mobile phones as an instrument for gathering *lifelog* information we have built a *Context Logger* application, which subscribes to all sensors through the relevant system components and then continuously records all events received from the multiple adapters and widgets. In effect all accessible sensor data is recorded, as shown in Table 1.

All sensor recordings are time stamped using the embedded mobile phone timer. The accelerometer was sampled every 2 seconds. Samples are read-out every 30s, in batches of 15 samples. The purpose of reading in bursts is to enable reading short bursts with higher sampling rate. GSM cellular information acquired included the Cell ID with country code, operator code, and location area code. In the present system the phone software API only allow reading the CellID of the GSM base transceiver station to which the phone is currently connected (not the ones visible). Since the GPS sensor is the most energy expensive sensor, it was only sampled 2–3 times an hour to reduce the energy consumption. Bluetooth scans was performed approximately every 1.5-3 minutes. The sampling rate varies as the Bluetooth discovery time increases with the number of Bluetooth devices available within discovery range. A discovery of a Bluetooth device will always produce the unique MAC address of the device, however, the lookup of the Bluetooth "friendly name" and device type might fail as more time is required to obtain this information. WLAN scanning is performed approximately once a minute, recording MAC address, SSID, and RX level (power ratio in decibels – a measurement of the link quality) of discovered Access Points. Finally, phone activity (SMS, MMS, and calls) were recorded whenever it occurred (phone number and direction).

In addition to the above mentioned sensor data, the *Context Logger* application allows a user to manually label his/her present location and activity. The label is a text string which can be entered by the user on the mobile phone, such as, *home*, *running*, and *having dinner*. Entered labels will be stored and subsequently shown on a list to pick from in order to avoid re-typing location and activity labels. Users can manually choose to label location and activity by

Sensor	Sampling	Data
Accelerometer	30/minute	3D Accelerometer values
GSM	1/minute	CellID of GSM base transceiver station
GPS	2–3/hour	Longitude, Latitude, and Altitude
Bluetooth	20–40/hour	Bluetooth MAC, friendly name, and device type
WLAN	1/minute	Access Point MAC address, SSID, and RX level
Phone activity	Event	Phone number and direction of call or message

Table 1. List of embedded mobile phone sensors used for collecting data

selecting a menu item in the *Context Logger* application, however, this mechanism is further enhanced with the ability to automatically prompt for labels. Thus, a user can choose to enter a label at any point in time, but the application will also prompt the user to label a location and activity 2–3 times a day in order to receive feedback. The data location and activity data recorded on the mobile phone is a key-value pair along with the time stamp. There are no pre-defined location and activity labels defined in the application and the labeling is completely free form with the users determining, how they want to write their text labels.

Situations with missing data may occur due to the phone running out of battery, being switched off, or simply not able to acquire data through one or more sensors (for instance no GSM reception).

2.3 Data Collection

An initial deployment of the system included continuous use by 14 participants, each equipped with a standard mobile phone (Nokia N95) which had the Mobile Context Toolbox pre-installed along with the *Context Logger* application that would continuously record the data acquired from all sensors currently supported by our framework. The participants were using the mobile phone as their regular mobile phone for a period of five weeks or more, as they were given instructions to insert their own simcard into the phone. Furthermore, they were instructed to make and receive calls, send messages, etc., as they would usually do, and generally use the phone as they would use their own phone. Therefore, no particular instructions were given, since we wanted to establish data from regular usage of the mobile phone, and thereby acquire real-life data. This means that the participants would not necessarily carry the phone on the body all the time such as carrying the mobile phone in a pocket.

As the survey is completely dependent on the cooperation of the participants and due to increased use of sensors, the lack of battery time was considered a risk in terms of participants leaving the survey. Thus, the sensor configuration of sampling on the phone was based on optimizing the resource consumption, so that the participants should only need to recharge the phone once a day (typically during the night).

The experiment started on October 28, 2008 and ended January 7, 2009 and the participants were students and staff members from The Technical University of Denmark volunteering to be part of the experiment. Thus, the participants were mainly situated in the greater Copenhagen area, Denmark. The 14 participants took part in the experiment between 31 to 71 days, resulting in approximately 472 days of data covering data collection periods totalling 676 days. The average duration was 48.2 days. An overview of the collected data is provided in Table 2.

During the experiment a total of approximately 20 million data points were collected with the accelerometer contributing the most with 14.5 million data points. A summary of recordings from Bluetooth scans, WLAN scans, GPS readings, and GSM readings can be seen in Table 2. It is worth noticing the

Part.	Accel	BT.	BT.*	GPS	GSM	GSM*	Ann.	PA.	WLAN	WLAN*	Days
1	1474480	54349	2846	516	69458	529	533	544	224101	6387	71
2	2045773	38028	2478	1514	75669	603	596	1062	364272	6040	66
3	318597	27329	790	12	37217	98	222	21	125600	630	31
4	875287	7880	743	2	17750	228	134	620	186421	2394	52
5	1117147	13575	2373	4058	56206	227	386	277	251016	2347	48
6	711490	23702	1141	95	51702	235	82	839	92396	2119	50
7	1184457	13327	1765	3	45826	955	272	581	139466	4017	46
8	700258	42346	2080	614	74250	172	212	74	154108	3359	41
9	1101926	42346	1050	119	37393	100	104	497	104576	1804	38
10	1103086	21676	2104	419	63937	419	414	116	192338	2650	48
11	1122315	12492	655	929	46158	929	163	121	295716	2286	47
12	796452	30610	2317	40	51548	40	143	151	97769	2403	50
13	1024276	27550	1741	1114	49349	1114	137	949	171951	5463	51
14	971558	21502	1303	44	40017	44	149	686	118687	1263	36
Total	14547102	350879	20408	9479	716480	2837	3547	6538	2518417	28110	48.2

Table 2. Overview of collected data for each participant in the experiment: Participant, Accelerometer, Bluetooth, Unique Bluetooth devices, GPS, GSM, Unique GSM cells, Annotations, Phone Activity, WLAN Access Points, Unique WLAN Access Points, Duration in days.

number of unique Bluetooth devices, unique GSM cells, and unique WLAN access points discovered accumulatively during the experiment by all participants: 20408, 2837, and 28110, respectively. In total 9479 readings of GPS position were recorded, but the recordings varies a lot among the participants due to the nature of the GPS technology. As a GPS position typically can not be obtained indoor only a subset of users have sufficient recordings of GPS position. For instance, if they typically place the mobile phone near a window when indoor where a GPS position can be obtained. A total of 6538 calls and messages took place during the experiment.

The participants provided 3547 annotations of locations and activities in total. On average the participants provided 253 annotations during their participation, with an overall average of 5.3 labels provided per user per day. The most active participants provided 8-9 labels per day on average, whereas the least active participants provided 2 labels per day on average.

3 Methods

In this study we will apply an information theoretic approach in the analysis of the dataset obtained and described in Section 2. Although before describing the details involved in this, we consider the preprocessing required for the final analysis.

A general issue in obtaining discrete times series is the number of quantization levels and sample rate of the true process [3]. The scan cycles used to obtain the present dataset are non-uniformly sampled and the scan cycles are of different length for each sensor. We therefore construct a commonly aligned time series for each sensor by creating non-overlapping frames of a given window length. The original samples falling within the frame is then assigned to it. If multiple samples falls within a frame they are merged, which is reasonable for the indicator type of sensors (WLAN, GSM, BT). This is similar to combining states in a Markov model effectively altering the state transitions. In case of the acceleration sensor, the feature is calculated as the average power within a window represented as a discrete levels[1], i.e., $X^{ACC} \in \{\text{off}, 1, 2, 3\}$. Considering the WLAN sensor and integrating all the networks seen into a state effectively means that the predictive variable becomes the WLAN state and not the location as such. If a specific location is needed a lookup to a database could return the position of the WLAN access point and generate a location variable from that. An alternative would be to directly work on a location variable generated from the WLAN access point, but this is not considered here.

The proposed representation constitutes a very detailed description of the participant state. An alternative suggested in [11, 12], represents the state as the most visited GSM cell location within a time window. Both approaches involve a relatively complex temporal quantization and resampling of the original data. To evaluate the consequence temporal scale, we consider the change in predictability bounds as the window length is decreased from one hour to one minute.

3.1 Information Theoretic Measures

We consider the problem of quantifying the predictability obtainable in a discrete process, $\mathbf{X} = (X_1, X_2, .., X_i)$, where i is the time index and X is the state variable. This is to a large degree motivated by previous work on predictability, complexity and learning (see e.g. [3]), and recent development on a similar dataset [11]. Here predictability is defined as the probability of an arbitrary algorithm correctly predicting the next state. Hence, given the history the basic distribution of interest is $P\left(X_{i+1}|X_1, X_2, .., X_i\right)$. In the case where we have no information regarding the history, the distribution naturally reduces to $P\left(X\right)$. When $P(X)$ is uniform, i.e., each of M states have the same probability of occurring, the Shannon entropy (in bits) is defined as $H^{rand}\left(X\right) = \log_2 M$.

In the case where the distribution of X is non-uniform, the entropy is given as

$$H^{unc}\left(X\right) = -\sum\nolimits_{i \in I} p\left(x_i\right) \log\left(p\left(x_i\right)\right) \qquad (1)$$

with $p\left(x\right) = P\left(X = x\right)$. This in turn represents the information when no history is available, hence, the acronym unc (uncorrelated).

The entropy rate of the participants trajectory, or the average number of bits needed to encode the states in the sequence, can be estimated taking into

[1] An equiprobable quantization is used, i.e., each level has the same frequency of occurrence within the entire dataset.

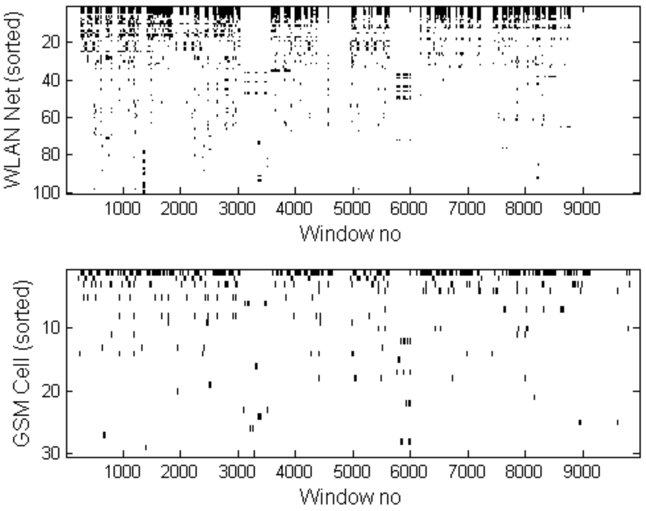


Fig. 2. Participant 2. Time series for WLAN (top 100) and GSM data (top 30). The time series are considered in a vector space representation, hence, each column is a state vector.

account the complete history. This is done by defining the stationary stochastic process $\mathbf{X} = \{X_i\}$ which have the entropy rate defined as

$$H\left(\mathbf{X}\right) = \lim_{n \to \infty} \frac{1}{n} \sum\nolimits_{i=1}^{n} H\left(X_i | h^{i-1}\right), \tag{2}$$

where the history h^i at time step i is $h^i = \{X_1, X_2, ..., X_{i-1}\}$. It is noted that $0 \leq H \leq H^{unc} \leq H^{rand} < \infty$.

A challenge in using these information measures based on real and unknown processes is the estimation of the entropy rate. A number of ideas have emerged based on compression techniques such as Lempel-Ziv (LZ) (including string matching methods) and Context Weighted Trees for binary processes. For a general overview, we refer to [6]. An appealing aspect of these non-parametric methods is that we avoid directly limiting model complexity as would be necessary if we applied parametric or semi-parametric models. In this study we estimate the entropy rate using a LZ based estimator as described in [8, 6] and also applied in [12]. The entropy rate estimate for a time series of length n is given by

$$H^{est} = \left[\frac{1}{n} \sum\nolimits_{i=1}^{n} \frac{L_i}{\log_2 n}\right]^{-1} \tag{3}$$

where L_i is the shortest substring starting at time step i, which has not been seen in the past. The consistency under stationary assumptions is proved in [8] where the method is applied to the analysis of English text.

Given estimates of the entropy and the entropy rate we consider a related quantity, namely mutual information - or predictive information [3]. This measure is already available given the information measures above as

$$I_{pred} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H\left(X_i\right) - H\left(X_i | h^{i-1}\right) \tag{4}$$

$$= H^{unc}\left(X\right) - H^{est}\left(\mathbf{X}\right) \tag{5}$$

In effect I_{pred} represents the mutual information between the distributions corresponding to knowing and not knowing the past history. Hence, it quantifies the fundamental information gain in knowing the (complete) past when prediction the future, and we propose it as a easy way to evaluate quantization and time scale effects. We illustrate the behavior of the measure on a time scale selection problem in Section 4.

3.2 Predictability

In order to construct bounds on the predictability we consider the probability, Π, that an arbitrary algorithm is able to predict the next state correctly.

Based on the entropy rate and Fano's inequality Song et al. derives a bound so $\Pi \leq \Pi^{\max}\left(H\left(X\right), M\right)$ with Π^{max} given by the relation [12]

$$H\left(X\right) = H\left(\Pi^{\max}\right) + \left(1 - \Pi^{\max}\right)\log_2\left(M - 1\right) \tag{6}$$

with the function, $H\left(\Pi^{\max}\right)$, given by

$$H\left(\Pi^{\max}\right) = -\Pi^{\max}\log_2\left(\Pi^{\max}\right) - \left(1 - \Pi^{\max}\right)\log_2\left(1 - \Pi^{\max}\right) \tag{7}$$

This non-linear relation between Π^{max} and the estimate of $H(X)$ is easily solved using standard methods (for a full derivation see [12]).

Adopting this approach we obtain three upper bounds based on the entropy estimates previously mentioned. The first, Π^{rand}, provides an upper bound on a random predictor. The second upper bound is Π^{unc} which bounds the performance obtainable with a predictor utilizing the observed state distribution. Finally, the most interesting bound, Π^{max}, provides a upper limit for the performance of any algorithm utilizing the complete past.

The upper bound is of course interesting in understanding the potential predictability, although we find a lower bound equally important in the analysis. Song et. al. [11] show how a simple lower bound can be constructed based on the so-called regularity. The regularity is in essence a zero order Markov model based on the most likely state at any given time of the day, i.e., using only the time of occurrence and no sequence information. This is an intuitive measure for some time periods such as daily patterns, e.g., utilizing where a person is most likely to be each morning at 7.00. However, if the time scale is in the minute range it does not necessarily make sense to consider the regularity.

Instead we propose to use a predictor using the immediate past as the representative of the lower predictability bound. For this purpose we use a first order

Markov model with the transition probabilities estimated from the finite process. Thus, the next state prediction is based on the distribution $P\left(X_{i+1}|X_1, X_1, ..., X_i\right) = P\left(X_{i+1}|X_i\right)$.

To avoid overfitting which tends to render the bounds overly optimistic, we use a resampling scheme in which the entropy and the next state distribution is estimated based on 2/3 of the data and tested on the remaining 1/3. This is performed for nine distinct subsections in a compromise between accuracy of the estimate and the needed samples for the entropy estimator to converge. The resampling further allows us to verify that the LZ entropy estimate converges to reasonable similar values for separate temporal sections of the participants life. Any variation across subsections will result in a greater variance of the estimated bound.

4 Results

In order to provide insight into the predictability of mobile phones sensors and thereby indirectly insight into human behavior, we apply the information theoretic methods described in Section 2. The density estimates of the bounds are all made using a standard kernel based density estimator (the bandwidth is hand-tuned for visualization).

4.1 Individual Sensors

One of the goals in this study is to analyse the potential predictability of different sensors and to that end we provide the predictability bounds for the four prominent sensors in the dataset, specifically GSM, WLAN, Bluetooth and acceleration at 15 min. window length. Fig. 3 shows the predictability bounds for the GSM, WLAN, Bluetooth and acceleration/activity sensors. By examining the difference between $\Pi^{\max}$ and Π^{unc} we find that there is considerable gain in knowing the past. From the difference between the Markov model Π^{Markov} and the upper bound, Π^{Max}, we see that knowing more than just the previous state seems to provide considerable benefit.

In the uncorrelated case, H^{unc}, participants show considerable differences in the entropy for all sensors as typically expected in a real population. However, conditioned on the past, the variability is typically lower (except for WLAN) indicating that participants with high entropy have a relative predictable trajectory. This corresponds well with the results observed in [11] for GSM based localization. The Bluetooth data does, as mentioned, contain a considerable fraction of unknown states (not off-states), which will tend to underestimate the true entropy. This means that the bounds are most likely overestimated for Bluetooth.

The GSM, WLAN and Bluetooth sensors are inherently location or proximity oriented sensors, and the estimated distributions all have modes in the high range above 80%, but with noticeable difference in variability. Whereas GSM provides small variability among participants, WLAN seems to provide a much larger difference among participants. This is not surprising since WLAN is considerably more noisy than GSM and captures a more local and detailed state than GSM. Thus, some participants tend to have a relatively low predictability while others are just as predictable as using GSM. The GSM complexity is on the other hand more similar between mobile phone users. In effect WLAN is most likely the more interesting sensor, but harder to predict, at least using the very detailed representation.

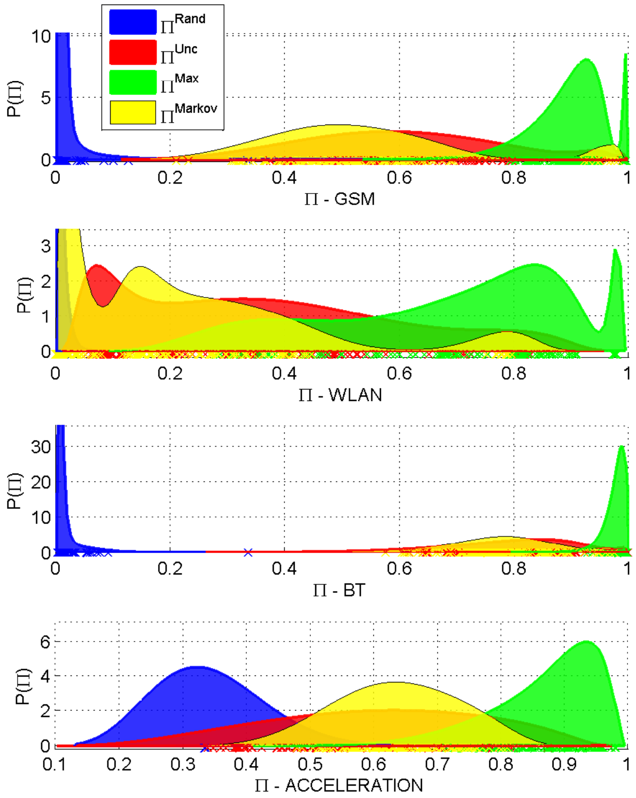


Fig. 3. Detailed representation: Predictability of GSM, WLAN, Bluetooth and Acceleration sensors. Crosses indicate individual participant estimates. The mode at 0.99 and 0.98 in the GSM and WLAN densities are due to participant 3 which is left out in the further analysis for clarity.

4.2 Time Scale

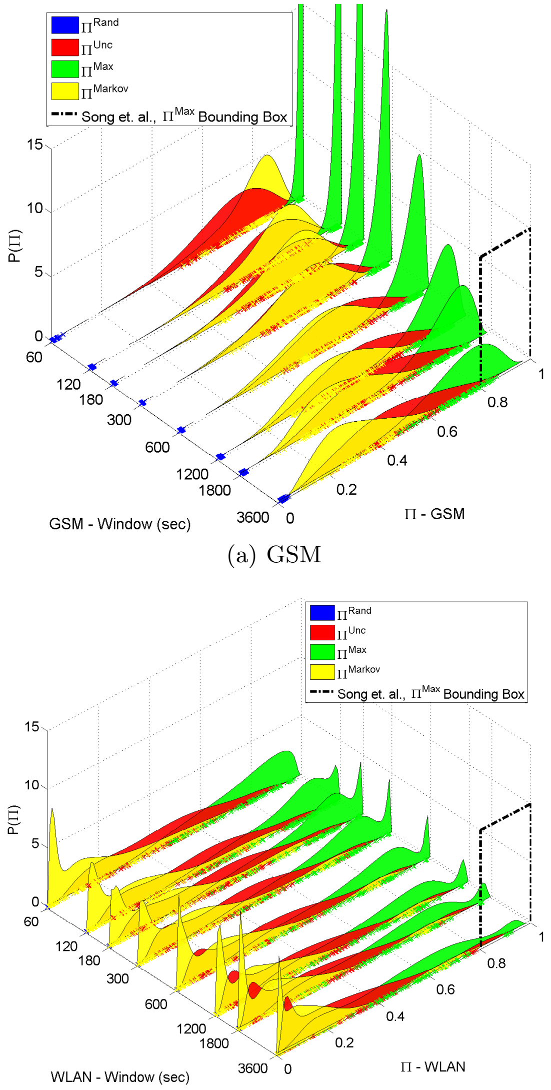
A primary goal in this study is the analysis of the time scales involved in the prediction of location based sensors with the aim to provide support for context-based services and general understanding of human mobility. The focus in this part is thus focused on the GMS and WLAN sensors and the predictability on a wide range of window lengths - and the examination of the optimal scale suggested by the predictive information.

Figure 4 shows how the predictability bounds changes with the window length. Noticeable are the GSM results in Figure 4(a) which are directly comparable with the original results in [11]. The bounding box indicates that the predictability is in the same range, although smaller in our case, possibly due to the more detailed representation utilized here. However, we obtain a similar upper bound at approximately 10 min. scale. This trend towards a high upper bound continues as the scale progresses downwards to 60 seconds.

In general there are various fundamental ways why this might happen. First of all, we may simply oversample a constant process leaving the resulting times series highly trivial to predict. The second reason is that the fundamental dependencies are removed when aggregating the cell at long scales and the shorter time scale provides the best representation. A simple way to examine the first options is to look at the predictability suggested by the first order Markov model and determine how far it is from reaching the upper bound. We notice that the despite Π^{Markov} approaching Π^{Max}, there seems to be some non-trivial behavior which is not predicted by the first order model. Not surprisingly this indicates that the first order Markov model is too simple. However, the bound provides a very convenient indication of what a more complex model is able to obtain.

Whereas GSM provides a rather rough indication of mobility and specific location, WLAN cells have quite high location resolution. Examining the time scale for WLAN reveals that the complexity of the problem is very high compared to the GMS case as seen on the pure results obtained by the Markov model. Despite this, we notice that the upper bound is still quite high. This suggests that there is an large unexploited potential in applying a more complex model than for example a first order Markov model. As with the GSM sensor we find that the shorter time scales provides a higher upper bound, and noticeably that the variability among the participants are lower, in effect offering better generalization of the predictability across multiple users.

As we have hinted, the optimal scale time scale for predictability for both GSM and WLAN at small time scales, and to examine the precise scale at which the past offers the most information in predicting the future we consider the predictive information. This is depicted in Figure 5 showing how the predictive information depends on the time scale. We find that the optimal scale is in the 3-4 minute range for both types of sensors. This is our main result and supplements the results in Song et al. [11] who focused on longer time scales (60 minutes). The high predictability at short time scales is of great interest for applications and is "good news" for pro-active services based on predicting human needs and behavior. Furthermore, the fact that the two distinct sensors operating on

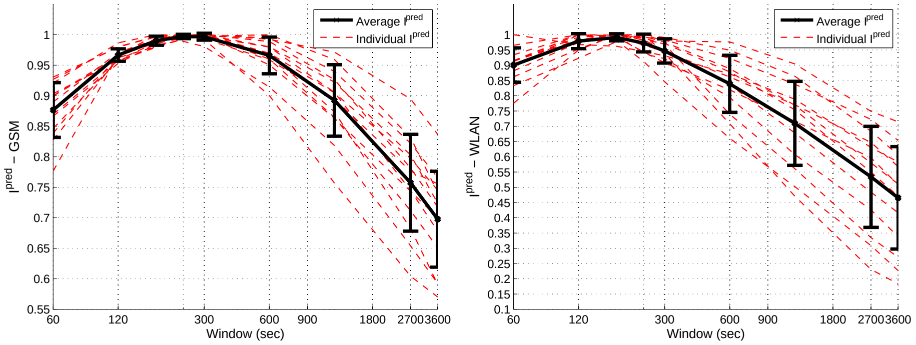


(a) GSM

(b) WLAN

Fig. 4. Predictability vs. window length in sec. (log scale). Notice that participant 3 has been removed from the density estimate due to his/her outlier nature as noticed in Figure 3

different spatial resolution yet still suggest the same optimal temporal scale, indicates that there exists fundamental information at this scale where both the upper bound on GSM and WLAN predictability are quite high. Yet, the exact information available at this scale and implications of this is to be examined in



(a) GSM

(b) WLAN

Fig. 5. Predictive Information (normalized) vs. window length (log scale). Participant 3 is left out.

a future analysis, for example using explicit modeling paradigms such as (multiway) factor analysis and advanced dynamical models.

5 Conclusion

In this paper we described an experimental setup for obtaining so-called *lifelog* data using embedded mobile phones. The resulting dataset offers many possibilities for investigating interesting elements of human behavior.

In the analysis we adopted an implicit modeling approach based on recent information theoretic methods to provide bounds for the prediction one could hope to obtain using explicit modeling. We presented results on the predictability of multiple mobile phone sensors showing that the basic findings in [11] generalizes to more location and proximity based sensors. Specifically, that the gain in knowing the past is significant, which indicates interesting potential for context-aware mobile applications relying on forecasting for example GSM and WLAN associations.

Finally, we showed that the prediction of human mobility generalizes to shorter time scales than the one hour time scale previously studied in [11]. In particular, we showed that the collected GSM and WLAN have the same optimal time scales for prediction, specifically 3-4 minute range. Despite this encouraging result, the exact interpretation and relevance of the patterns at the suggested scale needs further investigation and analysis, for example using explicit modeling.

References

1. MyLifeBits (accessed May 1st 2010). `http://research.microsoft.com/en-us/projects/mylifebits/default.aspx`

2. Bell, G., Gemmell, J.: A digital life. Scientific American 296(3), 5865 (March 2007)
3. Bialek, W., Nemenman, I., Tishby, N.: Predictability, complexity, and learning. Neural Comput. 13(11), 2409–2463 (2001)
4. Eagle, N., (Sandy) Pentland, A.: Reality mining: sensing complex social systems. Personal and Ubiquitous Computing 10(4), 255–268 (2006)
5. Farrahi, K., Gatica-Perez, D.: Discovering human routines from cell phone data with topic models. 2008 12th IEEE International Symposium on Wearable Computers pp. 29–32 (2008)
6. Gao, Y., Kontoyiannis, I., Bienenstock, E.: Estimating the entropy of binary time series: Methodology, some theory and a simulation study. Entropy 10(2), 71–99 (2008)
7. Jensen, B.S., Larsen, J.E., Jensen, K., Larsen, J., Hansen, L.K.: Estimating human predictability from mobile sensor data. In: IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010) (2010)
8. Kontoyiannis, I., Algoet, P., Suhov, Y.M., Wyner, A.: Nonparametric entropy estimation for stationary process and random fields, with applications to english text. IEEE Transactions on Information Theory 44(3), 1319–1327 (1998)
9. Kwok, R.: Personal technology: Phoning in data. Nature 458(7241), 959–961 (2009)
10. Larsen, J.E., Jensen, K.: Mobile context toolbox - an extensible context framework for s60 mobile phones. In: Proceedings of the 4th European Conference on Smart Sensing and Context (EuroSSC) (2009)
11. Song, C., Qu, Z., Blumm, N., Barabási, A.L.: Limits of predictability in human mobility. Science 327(5968), 1018–1021 (2010)
12. Song, C., Qu, Z., Blumm, N., Barabási, A.L.: Limits of predictability in human mobility - supplementary material (2010)