1

The Indoor Predictability of Human Mobility: Estimating Mobility with Smart Home Sensors

Tinghui Wang
School of EECS
Washington State University
Pullman, WA, USA
tinghui.wang@wsu.edu

Diane J. Cook, IEEE Fellow
School of EECS
Washington State University
Pullman, WA, USA
djcook@wsu.edu

Thomas R. Fischer, IEEE Fellow
School of EECS
Washington State University
Pullman, WA, USA
thomas_fischer@wsu.edu

Abstract—Analyzing human mobility patterns is valuable for understanding human behavior and providing location-anticipating services. In this work, we theoretically estimate the predictability of human movement for indoor settings, a problem that has not yet been tackled by the community. To validate the model, we utilize location data collected by ambient sensors in residential settings. The data support the model and allow us to contrast the predictability of various groups, including single-resident homes, homes with multiple residents, and homes with pets.

Index Terms—prediction, compression, human mobility, smart homes

I. Introduction

Studying human movement patterns helps researchers to understand human behavior. If human movement can be anticipated, this information can improve intelligent services that are provided by smart cities and smart homes [1]. One question that has captured the attention of researchers is determining how predictable human motion is. Given the ability to collect massive amounts of geolocation data, researchers have formalized models of large-scale outdoor human movement, population migration, and epidemics. These advances lead to improvements in city planning, traffic engineering, public health, and communication.

On the other hand, sensing indoor human mobility offers untapped potential to model behavior patterns. Indoor activities occupy most of our daily routine. According to the Environmental Protection Agency, Americans spend up to 89% of their time indoors, 69% specifically inside residences [2]–[4]. Using ambient sensors, researchers can continuously monitor the locations and activities of building occupants without requiring resident effort or behavior change [5], [6]. Discovering indoor mobility patterns facilitates providing mobility-aware services and automatic detection of abnormal behaviors. In buildings with multiple inhabitants, accurate prediction of each individual's movements is also key to tracking and responding to each person's needs.

In this work, we investigate the theoretical limit of the predictability of indoor human mobility. We also assess the performance of existing mobility prediction models. In this treatment, we refer to mobility prediction as the ability to predict the location of a human at a future point in time. Because we model human motion using smart home sensors, predictability is framed relative to this context. We therefore also consider the impact of smart home configurations, such as sensor resolution and the presence of multiple smart home residents, on mobility predictability.

We empirically validate our theoretical analysis by analyzing 140 million motion sensor readings recorded in 117 CASAS smart homes [7]. In these homes, ambient sensors including passive infrared (PIR) motion sensors, magnetic door sensors, and contact-based item sensors, are deployed to track residents and their interaction with objects in the home. The sensor map for a sample smart home site, home m3, is shown in Figure 1.

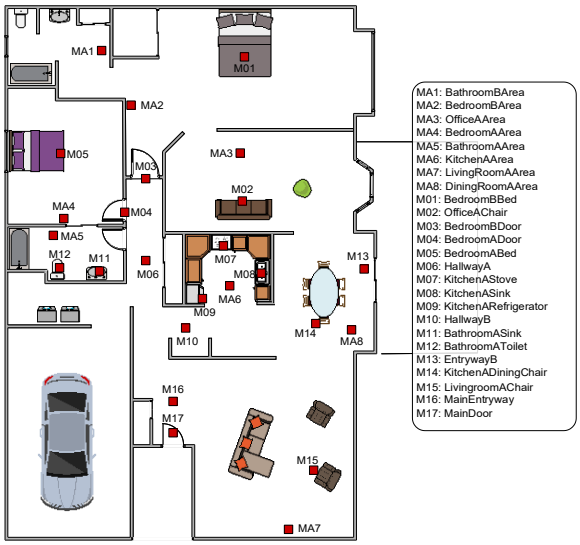

Fig. 1. Floor plan and sensor positions in a sample smart home (home m3).

Whenever the state of the sensor changes due to resident activity, the sensor sends a message to the smart home gateway, and the message is uploaded to a central database. Each sensor message is a three-tuple containing the time of the observation, the sensor identifier that observed the resident movement, and the state of the sensor. Table I shows an excerpt of sensor messages collected in one smart home testbed.

This material is based upon work supported by the National Science Foundation under Grant No. 1543656.

TABLE I
EXCERPT OF SENSOR EVENTS COLLECTED IN SMART HOME M3

Time Tag	Sensor ID	Message
2016/12/27 7:44:58	BedroomADoor	ON
2016/12/27 7:44:59	BedroomAArea	ON
2016/12/27 7:45:01	BedroomADoor	OFF
2016/12/27 7:45:05	BedroomAArea	OFF
2016/12/27 7:45:08	KitchenAArea	ON
2016/12/27 7:45:09	BedroomAArea	ON
2016/12/27 7:45:11	HallwayB	ON
2016/12/27 7:45:13	BedroomAArea	OFF
2016/12/27 7:45:14	BedroomAArea	ON
2016/12/27 7:45:15	BedroomAArea	OFF
2016/12/27 7:45:15	HallwayB	OFF

Because a "deactivate" message (OFF for motion sensors, CLOSE for magnetic door sensors, and PRESENT for item sensors) always follows an "activate" message (ON for motion sensor, OPEN for magnetic door sensor, and ABSENT for item sensors), we only analyze the "activate" messages. Data are continuously collected in these environments while residents perform their normal daily routines.

II. HUMAN MOBILITY

Human mobility has been the subject of investigation in many fields, but the advent of mobile computing has opened the door to creating statistical models based on passively-observed movement patterns. In their seminal work, Song et al. [8] formulated an upper bound on outdoor human mobility predictability by entropy rate estimation. In this study, geolocation data are obtained from cell tower logs. Here, an individual's location is identified by the closest mobile tower each time they use a phone. Data are then extrapolated to form a continuous hourly record for estimating the entropy rate. Based on Fano's inequality, the authors concluded that the potential predictability of human mobility could be as high as 93%. This early study ignited a series of follow-up work [9]–[14] that replicate the outdoor mobility experiments using data collected from high-resolution GPS, embedded sensors on mobile devices, and wireless signal strengths to understand mobility predictability at various spatial and temporal resolutions. However, Smith et al. [12] noticed that, in geolocation data, each individual only visits a small fraction of all locations. As a result, the upper bound claimed by Song et al. [8] is an overestimate, an observation that is echoed by Xu et al. [15], and the actual limits could be 11-24% lower.

A limitation of earlier investigations is that the geolocation data monitors only outdoor human movement. Understanding outdoor movement supports analysis of viruses, city planning, and management of resources. Unfortunately, the theoretical bound of indoor human mobility predictability is a subject that has not received this same level of investigation. Understanding indoor movement is also valuable. Researchers have used such information to detect changes in a person's health [16], identify possible in-home security threats [17], design proactive behavior interventions [18], and provide resident-aware home automation [19]. Predicting future resident locations is a central task of smart homes; thus researchers have proposed algorithms to model the indoor mobility of residents [20]. For example, Gopalratnam and Cook [21] proposed an online Lempel-Ziv sequence prediction algorithm, Active LeZi, which converges to 47% accuracy for predicting indoor locations using one month of smart home data. Alam et al. [22] introduced SPEED, an algorithm that predicts future events through a partial match, and reported 88.3% accuracy in predicting next sensor events over one month. Jayarajah and Misra [23] used WiFi logs on an urban campus to monitor indoor activities and observed an 87% predictability across building sections (e.g., classroom, food outlet, meeting space). Minor et al. [24] introduced a structured prediction algorithm that forecasted the times of upcoming activities with a mean absolute error (MAE) of 16 minutes, although this was not applied to movement prediction.

To analyze movement in smart homes with multiple residents, Ghasemi and Pouyan create a Markov chain that maps states to indoor sensors. This Markov model is used to predict the next location of each resident [25]. Doty et al. [26] further improve smart home state estimation by introducing flag hidden Markov models that contain a finite-state Markov chain as well as a structured observation process wherein a subset of states emit flags (i.e., observations) while other states are unmeasured. Lin and Cook [27] analyze the Markov order that best fits behavior observed by smart home sensors. While the focus of this paper is on deriving an upper bound on indoor mobility predictability rather than providing a best fit of smart home data, researchers including Begleiter et al. [28], Dimitrakakis [29], and Bejerano [30] have contributed several additional methods for improving prediction methods with variable-order models.

As an alternative method to model smart home behavior, Roy et al. [31] model the movement of smart home residents as actions taken in a multi-agent stochastic game. Thus, the mobility of all residents, characterized by a policy (i.e., the probability of a state-action pair), can be learned using a Nash H-learning algorithm. These authors reported 90% accurate prediction of a resident's room location over all time points for residents and 40% accuracy for visitors based on smart home data recorded over three weeks. These works are complementary to our analysis. However, they cannot be directly compared because we focus on the theoretical bound of indoor human movement predictability.

III. ENTROPY RATE AND PREDICTABILITY

The goal of this research is to investigate the theoretical bound for the predictability of indoor human mobility in real-world settings. Here we present the relationship between predictability and entropy rate estimated from actual smart home empirical data, using alternative computational approaches.

As introduced in Section I, the indoor locations of smart home residents are identified by ambient sensor messages. A person's indoor trajectory, T, is represented as a series of sensor messages, $\{X_i\}_{i \in Z}$, where X_i is the i^{th} sensor message in the trajectory. We define indoor mobility prediction as prediction of the next sensor message activated by the

smart home resident(s) based on their past trajectories. The resolution of the mobility prediction is thus impacted by the number of sensors in the environment and sensor resolution. In the case of multiple residents, we initially consider predicting the next sensor message caused by any of the residents at the site, without attempting to identify which of the residents was responsible for the message. Using this method, multiple-resident prediction can thus be viewed as prediction of a single, more "complex" resident in the building. The upper bound of the predictability of indoor mobility can be calculated by estimating the entropy rate of the underlying stationary stochastic process [8].

A. Entropy Rate and Predictability

We assume that ambient sensor-detected resident trajectory inside a building can be modeled by a stationary stochastic process $X = \{X_1, X_2, \ldots, X_n\}$, where X_i is a random variable representing the i^{th} sensor message. For smart home data, the values of each variable X_i are the identifiers of activated sensors, representing the location of a smart home resident at time i. The entropy rate, $H(X)$, measures the average conditional entropy $H(X_{n+1} \mid X_n, X_{n-1}, \ldots, X_1)$ when n approaches infinity, as shown in Equation 1. This rate can also be interpreted as the growth of the information contained in the trajectory compared to the length of the trajectory.

$$H(X) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_{i+1} \mid X_i, \ldots, X_1) \tag{1}$$

The most powerful prediction model for indoor mobil-ity, characterized by the conditional probability distribution $f(X_{n+1} \mid T_n)$, would predict the next sensor message (at time $n+1$) based on the complete past history of resident trajectories (for times 1 through n). Here, X_{n+1} represents the next sensor message, and T_n represents the past history of resident trajectory up to time step n. The maximum predictability Π^{max}, defined as the prediction accuracy of the next sensor message averaged over the entire observed sequence, can be formulated according to Equation 2. In this equation, $P(T_i)$ represents the probability at time i of having observed a particular resident trajectory T_i.

$$\Pi^{max} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{T_i} \sup_{X_{i+1}} f(X_{i+1} \mid T_i) P(T_i) \tag{2}$$

Given an observed trajectory T_i, a predictive estimator $g(T_i) = \hat{X}_{i+1}$ will output an estimated value of $\hat{X}_{i+1}$. The probability of the estimator's error is then defined as the probability that the estimate does not match the actual next sensor message, X_{i+1}, or $P_e(T_i) = P(\hat{X}_{i+1} = X_{i+1})$. Fano's inequality, which relates information lost in a noisy channel to the probability of predictive error [32], states that

$$H_b(P_e(T_i)) + P_e(T_i) \log_2(N) \geq H(X_{i+1} \mid T_i) \tag{3}$$

In Equation 3, $H_b(P)$ is the binary entropy function shown in Equation 4 and N represents the number of possible messages that could be observed (number of sensors in the home).

$$H_b(P) = -P \log_2 P - (1-P) \log_2 (1-P) \tag{4}$$

Given the conditional entropy $H(X_{i+1} \mid T_i)$, Fano's inequality guarantees a lower bound of the probability of error $P_e(T_i)$, and thus an upper bound of the predictability $\Pi(T_i) = 1 - P_e(T_i)$. We can represent the left side of Equation 3 as a function $\Pi(T_i)$, expressed in Equation 5.

$$H_F(\Pi(T_i)) = H_b(1 - \Pi(T_i)) + (1 - \Pi(T_i)) \log_2 N \tag{5}$$

Based on the concavity of $H_F(\Pi(T_i))$, we can associate the maximum predictability Π^{max} with the entropy rate of the underlying stationary stochastic process by applying Jensen's inequality, as shown in Equation 9.

$$H(X) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_{i+1} \mid X_i, \ldots, X_1) \tag{6}$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{T_i} H(X_{i+1} \mid T_i) P(T_i) \tag{7}$$

$$\leq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{T_i} H_F(\Pi(T_i)) P(T_i) \tag{8}$$

$$\leq H_F \left(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{T_i} \Pi(T_i) P(T_i) \right) \tag{9}$$

$$= H_F(\Pi^{max}) \tag{10}$$

According to Equation 10, given the entropy rate of the underlying stochastic process, the upper bound of predictability of the sequence can be calculated numerically.

B. Estimating Entropy Rate

Having established the association between entropy rate and predictability, the key challenge is to reliably estimate the entropy rate from empirical data with a universal coding method. The entropy rate, defined in Equation 1, is the conditional entropy of the future random variable as the length of the random process approaches infinity. Due to the limited sample size of our empirical data, we implement multiple entropy rate estimators to ensure the consistency of entropy rate estimation.

Previous work by Song et al. [8] and Smith et al. [12] used an entropy rate estimator based on Limpel-Ziv data compression, proposed by Kontoyiannis et al. [33]. In this approach, let $X = \{X_i\}$ be a stationary ergodic process with entropy rate $H(X) > 0$. We can then state:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\Lambda_i^i}{\log_2 n} = \frac{1}{H(X)} \tag{11}$$

In Equation 11, Λ_i^k is the length of the shortest substring starting at position i that does not appear as a continuous substring of the previous k symbols.

The entropy rate can also be estimated using the code length after compressing the data. Let $R(T_n)$ denote the size in bits of a sequence $T_n = \{X_n, X_{n-1}, \ldots, X_1\}$ after performing data compression. Takahira et al. verify that the code length per unit, $r(n) = \frac{1}{n}R(T_n)$, is always larger than the entropy rate $H(X)$ [34]. Additionally, provided the stochastic process is stationary and ergodic, a universal text compressor guarantees that the encoding rate converges to the entropy rate. Thus, we estimate the entropy rate by calculating data compression. In our experiments, we compare several state-of-the-art data compressors for this process. Specifically, we select 7-zip deflate (an implementation of Lempel-Ziv 77 [35] with Huffman coding), 7-zip LZMA (an optimized version of LZ77), and 7-zip PPMD (a lossless data compression using prediction by partial matching [36]).

Though all of the above approaches will converge to the true entropy rate, the accuracies of the estimates differ and are affected by the data sample size. To demonstrate the differences between the estimators and contrast the estimated values with the true entropy rate, we generated a state sequence of 10,000,000 time steps based on a second-order Markov chain with 20 states, where the probability of the current state is dependent on the previous 2 states. The state transition probabilities of the second-order Markov chain, characterized by $P(X_{i+1} \mid X_i, X_{i-1})$, are generated randomly. Thus, the true entropy rate of the synthetic state sequence (i.e., the entropy of the simulated second-order Markov chain) can be calculated based on the entropy rate definition shown in Equation 12.

$$\begin{aligned} H(X) &= H(X_{i+1} \mid X_i, X_{i-1}) \\ &= \sum_{x_i, x_{i-1}} P(x_i, x_{i-1}) H(X_{i+1} \mid x_i, x_{i-1}) \end{aligned} \tag{12}$$

In Equation 12, $P(x_i, x_{i-1})$ is the stationary probability distribution of state sequence $\{x_i, x_{i-1}\}$, and $H(X_{i+1} \mid x_i, x_{i-1})$ is the entropy of random variable X_{i+1} given the values of the previous two states.

In Figure 2 (top), we plot the entropy rates based on sequence size. For this plot we compare multiple established methods including the LZ77 deflate (H^{LZ77}), LZMA (H^{LZMA}) and PPMD (H^{PPMD}) algorithms, as well as the estimator defined in Equation 11 (H^{est}). Figure 2 (bottom) plots the corresponding predictability of each estimator. To put the results in perspective, we also include the random entropy (H^{rand}), temporally-uncorrelated entropy (H^{unc}), and entropy rate (H^{MC}), of a first-order Markov chain fitted to the synthetic sequence. The random entropy $H^{rand} = \log_2 N$ represents the maximum amount of possible information content in any sequence of N states. The uncorrelated entropy H^{unc} assumes that there is no temporal correlation between consecutive random variables in the random process. In other words, the sensor message at each time step is independently drawn from a probability distribution $P(X_i)$. Thus, H^{unc} can be calculated according to Equation 13, where $P(X_i)$ can be estimated by counting the occurrences of each observable symbol.

$$H^{unc} = -\sum_{x_i} P(X_i) \log_2 P(X_i) \tag{13}$$

The Markov chain-based entropy rate, H^{MC}, assumes that the state sequence can be modeled by a Markov chain, characterized by the conditional probability of state transition $P(X_{i+1} \mid X_i)$. Based on the empirical data, the transition probability can be estimated by maximizing the likelihood of the observed sequence. Thus, the entropy rate of the constructed Markov chain can be calculated according to Equation 14 [32]. Here, $P(X_i)$ is the stationary state distribution of the Markov chain, calculated by solving Equation 15.

$$H^{MC} = -\sum_{X_{i+1}, X_i} P(X_i) P(X_{i+1} \mid X_i) \log_2 P(X_{i+1} \mid X_i) \tag{14}$$

$$P(X_i) = \sum_{x_j} P(X_i \mid X_j) \quad \forall j \tag{15}$$

In Figure 2, the true entropy rate (the red dashed line) is calculated according to Equation 12. As shown, all estimators (H^{LZ77}, H^{LZMA}, H^{PPMD}, and H^{est}) converge to the true entropy rate as the sequence length increases. Among the estimators based on the compressed code length (H^{LZ77}, H^{LZMA}, and H^{PPMD}), PPMD yields the tightest upper bounds and fastest convergence. Due to the dictionary size, the entropy rates reported by those estimators are higher than the actual process entropy rate. Comparatively, H^{est} provides the most accurate estimate among all of the tested entropy rate estimators. While the estimators based on compress code length are guaranteed to be an upper bound to the actual entropy rate, H^{est} may undershoot when there is not enough data, leading to an over-estimated predictability [37]. The estimators based on compressed code length are guaranteed to be an upper bound to the actual entropy rate.

The entropy rate of any stochastic process with an alphabet size of 20 should lie between the random entropy H^{rand} (4.33 in this example), indicating that the process is entirely random, and 0, indicating that the process is fully deterministic. The second-order Markov model has an actual entropy rate of 1.35, and thus a corresponding predictability of 83.51%. The entropy rates estimated by H^{est} and H^{PPMD} are 1.52 and 1.60, resulting in an error of predictability of 3.5% and 4.01%, respectively. Since the synthetic data in this example is generated by a second-order Markov chain, a first-order Markov chain is not sufficient to capture patterns in the data. As a result, entropy rate H^{MC} calculated by fitting a first-order Markov chain to the synthetic data is 3.58, much higher than the actual entropy rate of the second-order Markov chain or the estimates reported by H^{est} and H^{PPMD}.

Additionally, the predictability estimated by H^{MC} is more than 30% lower than estimates derived from H^{est} or H^{PPMD}. Compared with H^{MC}, the temporally-uncorrelated entropy H^{unc} is less powerful for capturing regularities in the synthetic data. Quantitatively, the predictability corresponding to the

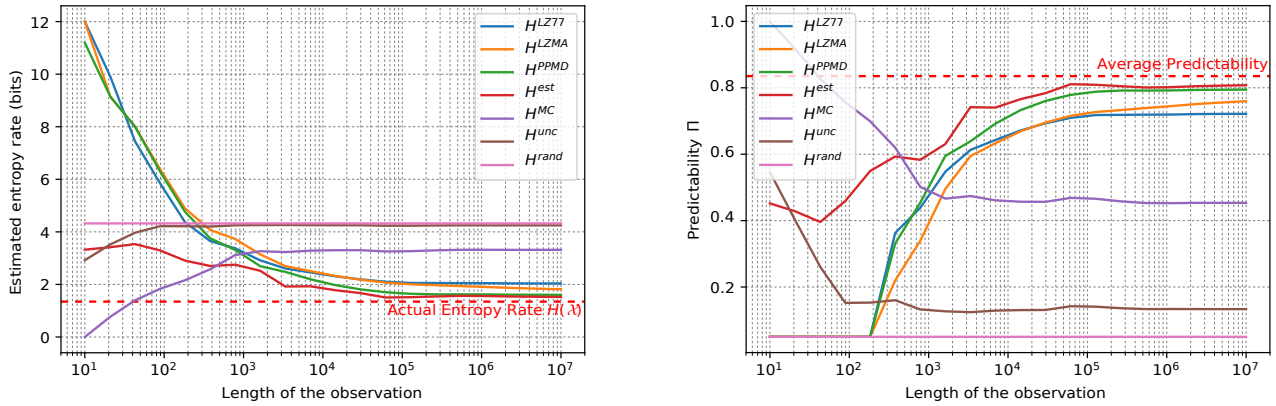

Fig. 2. Entropy rate (left) and predictability (right) estimated from synthetic data generated by a second-order Markov chain.

uncorrelated entropy is close to a fully-random sequence, which is more than 40% lower than the first-order Markov chain, and about 80% lower than the predictability calculated from the actual entropy of the sequence. In this example, the random entropy $H^{rand} = 4.33$ sets the upper bound of the information contained in any data sequence with 20 symbols (the number of states in this case). The random entropy can only be achieved if, at any time step, each state is equally probable, meaning $P(X_i) = 1/20$. Thus, the average predictability of such a random stochastic process equals 0.05, as shown in Figure 2.

IV. Predictability of Indoor Human Mobility

In the previous section, we investigated a formal method to determine the theoretical limits of indoor human mobility predictability in real-life settings. Here, we validate our theoretical formulations based on sensor message data collected from 117 smart homes, characterized in Table II. We collected these data from participants using the CASAS Smart Home in a Box technology [7], [38]. Residents in the smart homes performed their normal daily routines while ambient sensors collected data. Data collection in these homes ranged from 2 weeks to over 11 years. Among these homes, 45 smart homes have ≥ 2 residents and 14 of these include pets (e.g., cats, dogs). The remainder of the 72 smart homes are single-resident, 10 of which include pets. Each room (or functional area) typically contains 2-3 sensors. The total number of sensors ranges from 14 to 70 (mean=26.61, std=8.93). Many of these datasets, as well as the code, are available online[1].

Since the estimated entropy rate accuracy is affected by the data sample size, we start by establishing the relationship between entropy rate estimators and the length of the recorded message sequence. For all of the evaluations, the estimators are trained separately from a random initialization for each home. Figure 3 plots the estimated entropy rates (top) and the corresponding predictability (bottom) against the number of sensor messages. In this example, the sensor messages were collected in a home with multiple residents, home m3, over 6 months. This home was chosen because it houses multiple residents (making the data complex). Furthermore, ground truth resident labels are provided for a subset of this home's data, associating each sensor message with a corresponding resident ID. To provide a time reference in the graphs, the plots include red vertical lines indicate the first, second, third, and fourth weeks, as well as the second, third, and sixth months, on a log scale. Entropy rates in Figure 3 are estimated using the code length of the compressed data based on the LZ77 (deflate) (H^{LZ77}), LZMA (H^{LZMA}), and PPMD (H^{PPMD}) algorithms, as well as the entropy rate estimator H^{est} defined in Equation 11. The results are compared against random entropy H^{rand}, uncorrelated entropy H^{unc}, and the entropy rate H^{MC} calculated by fitting a first-order Markov chain to the sensor messages. The predictabilities plotted in Figure 3 are calculated according to Equation 10. Since there are 25 sensors in home m3, the size of alphabet N in Equation 10 is 25.

Among the entropy rate estimators (H^{LZ77}, H^{LZMA}, H^{PPMD} and H^{est}), H^{est} provides the lowest estimates. In the figure, H^{est} oscillates when there is not enough data, and stabilizes after consuming two months of data. Based on such observations, we acknowledge that the entropy rate reported by H^{est} is the best approximation to the true entropy rate of the underlying stochastic process that generates the observed human trajectories. As shown in Figure 3, the upper bound of indoor mobility predictions in this home converges to 79.3% according to the entropy rate estimated by H^{est}. If a first-order Markov model is used to predict the next sensor message triggered by the residents living in the smart home, the prediction accuracy, according to the entropy rate H^{MC}, is 67.5%, indicating that an improvement of 11.8% can potentially be achieved.

By repeating the above experiment for each smart home, we calculated the predictability upper bound of each home. Figure 4 summarizes the upper bounds of resident mobility predictability using entropy estimator H^{est}. Homes with single

[1]Links to code and datasets are online at http://casas.wsu.edu

TABLE II
SUMMARY OF THE SMART HOME DATASETS USED FOR EVALUATION

Site	Sensors	Residents	Size (sf)	Messages	Weeks	Site	Sensors	Residents	Size (sf)	Messages	Weeks
m1	27	2	1400	1,178,506	54	m2	42	2	1600	6,826,679	236
m3	25	2-3	1803	1,378,574	31	m4	70	2	1106	18,944,701	614
m5	36	2-3	1106	973,349	52	m6	19	2	1000	1,155,121	31
m7	22	2	1200	1,550,683	132	m8	17	2-3	3000	980,093	10
m9	39	2-3	1765	409,115	23	m10	47	2-3	4032	176,412	8
m11	22	2-3	3212	2,144	2	m12	37	2-3	2144	129,213	6
m13	35	2-3	1759	60,377	13	m14	60	2-3	2955	426,996	13
m15	40	2-3	2898	236,637	8	m16	51	2-3	3323	293,211	10
m17	24	2	N/A	153,603	10	m18	27	2	N/A	228,634	10
m19	22	2	N/A	376,708	10	m20	30	2	N/A	840,649	16
m21	30	2	N/A	377,391	13	m22	35	2	1400	1,432,718	80
m23	36	2	N/A	70,066	3	m24	27	2	N/A	249,987	10
m25	19	2	N/A	307,187	10	m26	25	2	N/A	284,137	11
m27	36	2	N/A	220,960	14	m28	24	2	N/A	324,905	12
m29	34	3	N/A	472,826	15	m30	30	2	N/A	114,684	4
m31	30	2	N/A	249,387	14	b1	27	2	1300	328,260	9
b2	33	2	1200	437,733	23	b3	25	3	N/A	784,428	21
b4	20	2	N/A	120,469	4	b5	27	4	N/A	594,493	13
b6	19	2	N/A	423,214	14	b7	21	2	N/A	242,491	10
b8	33	2	N/A	311,322	15	b9	21	2	N/A	288,301	13
b10	30	3	N/A	1,431,903	24	b11	30	3	N/A	599,925	19
b12	30	2	N/A	249,289	11	b13	32	2	N/A	528,142	15
b14	35	2	N/A	63,146	6	p1	20	1	800	1,509,615	82
p2	31	1	1100	207,690	12	p3	29	1	N/A	224,022	10
p4	18	1	800	759,750	51	p5	20	1	1000	3,884,561	220
p6	26	1	N/A	611,724	18	p7	30	1	N/A	296,860	14
p8	23	1	N/A	286,749	18	p9	18	1	N/A	489,311	21
p10	35	1	N/A	91,890	7	s1	32	1	600	327,271	26
s2	32	1	600	285,207	26	s3	32	1	600	256,301	26
s4	19	1	700	639,517	54	s5	20	1	700	56,729	7
s6	34	1	1400	2,583,713	92	s7	30	1	1500	2,399,519	146
s8	21	1	800	2,683,317	299	s9	33	1	1100	2,460,156	118
s10	30	1	800	3,154,735	276	s11	39	1	1200	2,028,153	169
s12	33	1	900	3,395,594	269	s13	27	1	800	5,393,455	239
s14	33	1	1100	3,423,249	309	s15	20	1	700	298,254	21
s16	29	1	800	1,117,701	73	s17	26	1	600	3,778,629	273
s18	20	1	800	818,002	51	s19	19	1	800	205,246	24
s20	16	1	700	385,107	44	s21	25	1	1200	2,991,898	153
s22	27	1	700	1,528,537	223	s23	20	1	700	906,345	100
s24	28	1	1100	438,681	31	s25	16	1	700	845,371	85
s26	18	1	700	12,299	22	s27	32	1	700	2,460,714	220
s28	17	1	700	1,654,459	178	s29	17	1	800	1,161,989	181
s30	20	1	900	4,323,961	175	s31	20	1	600	1,715,885	98
s32	26	1	700	371,175	27	s33	14	1	800	1,379,134	211
s34	20	1	900	1,155,121	211	s35	15	1	600	1,681,679	207
s36	18	1	N/A	217,944	8	s37	25	1	N/A	116,878	14
s38	24	1	N/A	280,500	12	s39	32	1	N/A	267,881	12
s40	20	1	N/A	86,996	10	s41	22	1	N/A	364,617	9
s42	28	1	N/A	287,873	14	s43	23	1	N/A	565,153	20
s44	17	1	593	888,745	60	s45	17	1	593	2,745,749	28
s46	19	1	837	5,321,015	61	s47	19	1	834	9,768,315	90
s48	17	1	593	773,010	67	s49	19	1	665	439,836	43
s50	22	1	1551	395,083	20	s51	30	1	1816	596,031	20
s52	17	1	N/A	316,837	20	s53	21	1	N/A	185,086	19
s54	20	1	N/A	517,534	15	s55	26	1	N/A	285,168	10
s56	31	1	N/A	311,815	12	s57	29	1	N/A	254,952	10
s58	30	1	N/A	700,171	22	s59	23	1	N/A	611,557	19
s60	36	1	N/A	468,679	10	s61	36	1	N/A	134,012	11
s62	19	1	N/A	117,986	12						

residents are shown in red and multi-resident homes are colored blue. Homes with pets as well as human residents are indicated with a star symbol. The indoor mobility of residents in some of these homes exhibits high orders of regularity, where the prediction accuracy of resident movement exceeds 95%. As expected, these occur in single-resident homes with no pets. On the contrary, in many multi-resident smart homes, the predictability is much lower and, in some cases, drops below 65%.

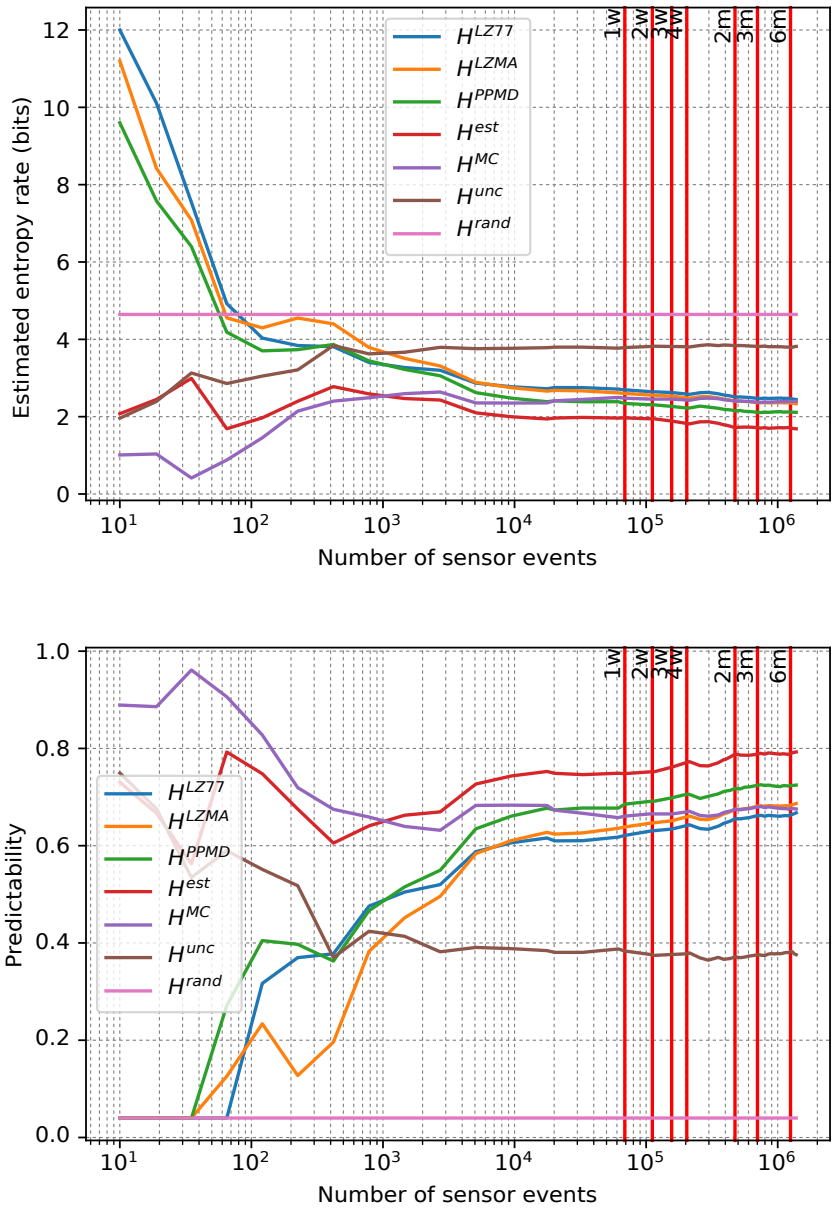

Fig. 3. Entropy rate (top) and corresponding upper bounds of predictability (bottom) estimated based on sensor messages recorded in a sample smart home (home m3).

A. Homes with Multiple Residents

To further explore the effect of multiple residents and pets on indoor mobility predictability, we pictorially compare predictability between a single resident smart home and a multi-resident smart home. Figure 5 shows two resident trajectories that are subsets of the collected sensor data. One trajectory is recorded in a single-resident smart home (home s4) and the other is recorded in a multi-resident smart home (home m3). In the single-resident trajectory recorded in home s1, the resident generally triggers sensors along the path as the individual moves from one part of the home to another. In this case, the next sensor message is expected to be strongly correlated to the previous sensor message. In multi-resident home m3, the merged (dual-resident) trajectory is indicated by red arrows. The actual paths of each resident, identified as R1 and R2, are shown with dotted green arrows and dotted blue arrows, respectively. In multi-resident smart homes, given that sensor messages are not mapped to specific residents, the combined movement is more complex and, therefore, more difficult to predict. As a result, we expect that the predictability of resident

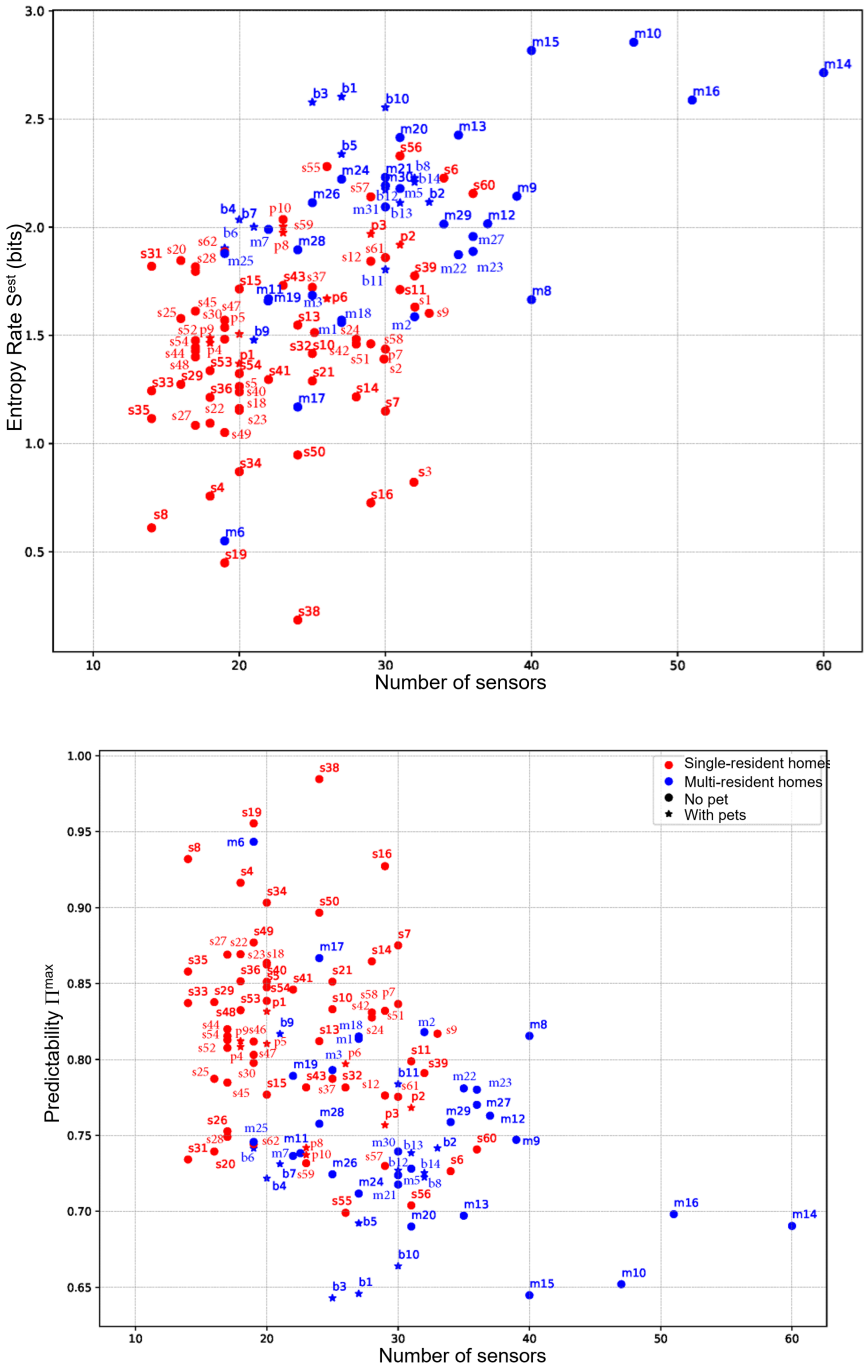

Fig. 4. Scatter plot of the entropy (top) and predictability (bottom) of resident mobility estimated in each smart home as a function of the number of sensors in the home. Single resident smart homes are represented as red circles, and multi-resident smart homes are represented in blue. A star marker indicates that the household includes pets.

mobility is lower in multi-resident settings when compared to the single-resident scenario. According to the statistics shown in Figure 6, the predictability limit of resident mobility in single-resident homes averages 81.86%, while the upper bound of mobility predictability in multi-resident settings is 74.40%, approximately 7.46% lower on average.

In addition to the uncertainty caused by multiple residents, pets in the household can also trigger ambient sensors and cause the creation of sensor messages. Because of their smaller mass, some pets do not consistently trigger passive infrared motion sensors, causing what appears to be a “teleporting” effect. This causes an increase in data noise and an overall decrease in predictability. In the experiments, we studied the predictability limits of smart homes with pets in comparison with the smart homes without pets. Based on these results, we found that the decrease in the predictability of indoor

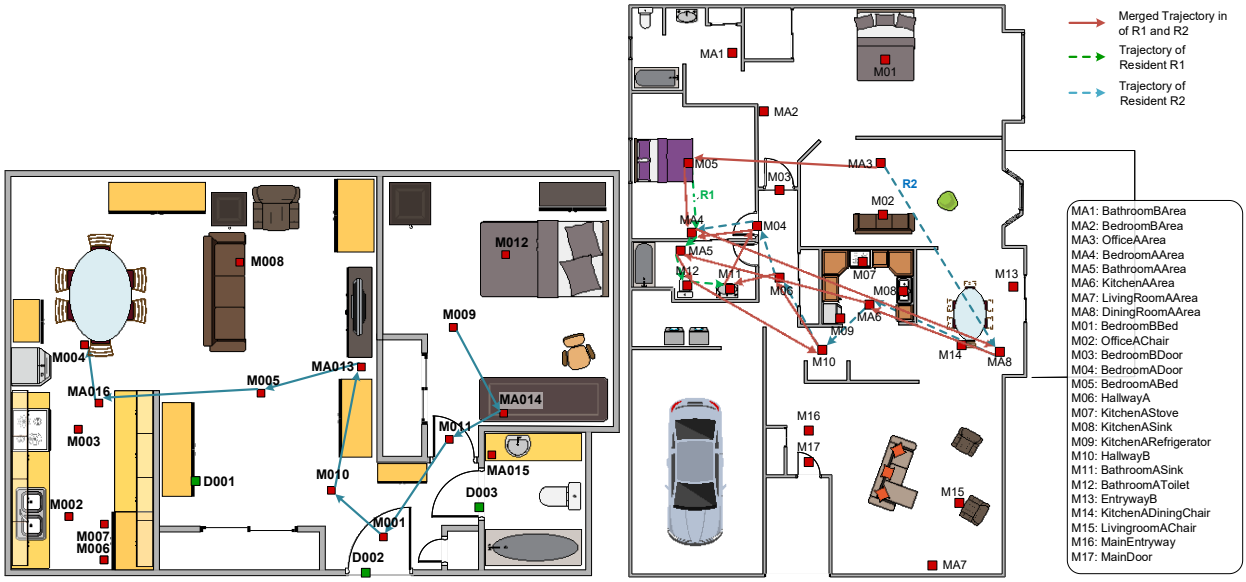

Fig. 5. Example of resident trajectories observed in a single-resident smart home (home s4, left) and a multi-resident smart home (home m3, right).

mobility caused by pets is 3% in both single-resident and multi-resident scenarios. Specifically, the predictability limit for single resident homes with pets averages 79.00% and multi-resident homes with pets averages 72.10%.

Because of the inherent difficulty in associating sensor messages with specific individuals in multi-resident scenarios, a simple Markov chain is commonly used to construct a human mobility model based on recorded sensor messages for these situations [25]. A Markov chain conditions the probability distribution of the next sensor message only on the previous sensor message. To model multi-resident movement, we initially fit a first-order Markov chain to the recorded data and calculated both entropy rates and associated predictability. In Figure 6, we average these values across all single-resident homes without pets, multi-resident homes without pets, single-resident homes with pets, and multi-resident homes with pets. We also compute and plot the theoretical upper bound on predictability for these house categories. In single-resident cases where there are no pets in the household, we found that the predictability of the first-order Markov chain is 6% below the theoretical upper bound of the estimated predictability, while in multi-resident homes or smart homes with pets, the predictability is closer to 8% below the theoretical limit. We note that the multi-resident homes exhibit the greatest variability in number of residents as well as overall movement patterns. At any particular point in time, the number of residents in these spaces can vary from 0 to 4+. As the number of residents increases, so may the randomness observed in the sequence. Based on the these results, there is certainly room for improvement in constructing more representative mobility models than are found in first-order Markov chains, especially in multi-resident settings.

One possible improvement to consider is employing methods to track multiple resident movements within a smart home and use this information to disaggregate the combined sensor data time series into separate components. If such a method can effectively track each resident, then we can analyze each resident separately. We explore this idea by utilizing GAMUT, a smart home multi-resident tracking algorithm [39]. GAMUT associates each sensor reading with a corresponding resident. To do this, the algorithm maps sensors onto a latent space using a technique borrowed from word embeddings [40]. GAMUT maps sensors to a latent space such that the distance between a pair of sensors in that space reflects how often one sensor in the pair generates a reading soon after the other.

To track multiple residents based on observed sensor readings, a Gaussian mixture probability density filter models each resident's likely movements. The model needs to account for possible movements within the home as well as cases when a resident enters or leaves, false readings due to noise, and reading caused by more than one resident moving together. Ultimately, a sensor reading is assigned to a resident that yields the highest likelihood. Based on the probability hypothesis density of the residents through time, GAMUT also estimates the number of residents that are in the space at any given time. From this information, one time series containing sensor readings corresponding to movements for r residents is automatically split into r time series, one per resident. This method yielded association accuracies of 0.82 for a home with 2 residents and 0.80 for a home with 4 residents [39]. In these homes, external staff manually annotated 12 days of sensor readings with corresponding resident identifiers to provide ground truth labels. The results of this analysis are included in Figure 6 and show that neither the estimated predictability nor the Markov-modeled predictability exhibit a large deviation from the original approach of modeling multiple residents as a single complex entity.

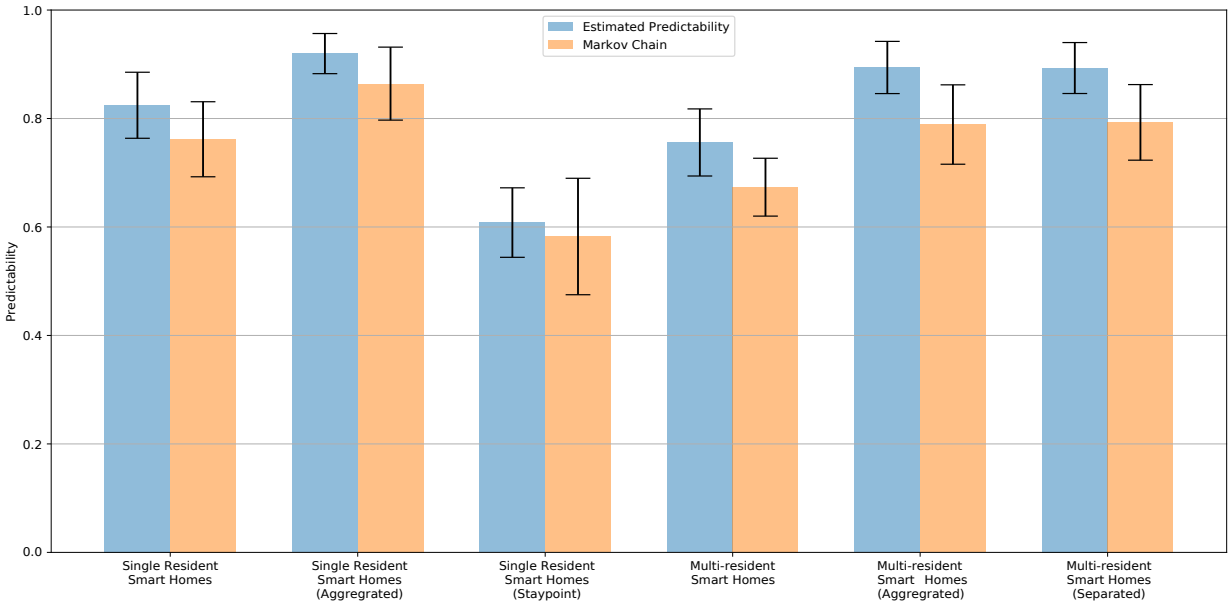

Fig. 6. Comparison of resident mobility predictability scenarios using Equation 11 versus a first-order Markov chain estimation (see Section IV-A). The scenarios include all sensor readings in single or multi-resident homes at the original sensor resolution, all sensor readings in single or multi-resident homes at a room-level sensor resolution (aggregated, see Section IV-C), staypoint-only sensor readings for single-resident homes at the original sensor resolution (Staypoint, see Section IV-B), or all sensor readings in multi-resident homes at the original sensor resolution where data are analyzed separately for each resident (Separated, see Section IV-A).

B. Staypoints

Our analysis so far focused on all movements that a resident makes within a home. A separate but related consideration is the predictability of a resident's intended location destinations. For example, if the resident is heading from the kitchen to an office area, sensors will generate readings along the path between the origin and destination locations. These intermediate locations may be more geometrically constrained (e.g., sensors in doorways and along hallways) and thus boost predictability.

To assess intended location predictability, we filter the original data to only consider "staypoints", or sensor-observed regions where the resident stays for at least one minute before moving to another location. The results for staypoint analysis of single-resident homes is shown in Figure 6. As the graph indicates, predictability is lower for staypoint locations than for all locations. The drop in predictability occurs for both the estimated predictability and the predictability of the Markov model.

C. Sensor Resolution

In this section, we analyze the observed impact of varying spatio-temporal sensor resolution on movement predictability. Resolution differences are very apparent when comparing indoor and outdoor mobility. A notable difference between outdoor and indoor mobility tracking is the data resolution. Outdoor analyses based their findings on hourly reports with spatial resolutions varying from 10 meters [9] to 3 kilometers [8]. In the case of our indoor analysis, the mobile sensors update their state every 1.25 seconds and the spatial resolution is 1 meter. Lin et al. [9] note that a trade-off exists between predictability and spatial resolution, thus the observed difference in resolutions indicates that indoor predictability will likely be higher than outdoor predictability. Another factor in comparing these sources of information is that the smart home sensors provide constant monitoring. In contrast, Lin et al. and Song report up to 25% and 80% missing locations, respectively. Such missing information introduces uncertainty into the empirical estimates.

One consideration for indoor mobility is the size of the environment. In our sample, this did not have a dramatic impact. Considering only single-resident homes, the mean predictability for the six largest (reported) homes was 0.833 (sd=0.053). This value is very close to the mean predictability for the six smallest homes of 0.828 (sd=0.025). Because all homes contained 2-3 sensors per functional area, the smaller homes had a higher density of sensors in the space. Thus, we focus our attention in this section on the impact of sensor resolution on mobility prediction.

To provide insight on the relationship between spatial resolution and predictability, we simulate a lower spatial resolution by aggregating individual sensors and their readings into just one "simulated" sensor for each room or region of the home. On average, we are replacing three sensors in the original data with one sensor in the aggregated data. The results are shown in Figure 6. As the graph indicates, the predictability of both single resident homes and multiple resident homes increases as the sensor resolution decreases. In these cases, there are fewer states to model and predict, thus simplifying the prediction problem.

D. Sensor Reliability

We note that the empirical results discussed here are subject to sensor error. For PIR motion sensors, error can originate from multiple sources. In some homes, there may be gaps in coverage. As a result, residents may move to locations that are not reflected in the sensor message sequence. Occasionally, motion sensors can generate false positive messages when hit with heat from an outside source such as a laser printer or baseboard heater. We estimated these errors occurred fewer than 0.05% of the days that were monitored. More commonly, sensor messages may be lost due to communication errors. In the CASAS smart homes, these are alerted as "radio errors." Such errors occurred on less than 0.45% of the monitored days.

Some of the errors due to ambient sensors could be corrected by fusing multiple sensor sources [41], such as fusing ambient sensor data with that of wearable and object sensors. Each data source is faced with challenges including sen-sor noise, participant non-compliance, and gaps in coverage. However, fusing data from multiple sources can harness the strengths of the individual sensor modalities to compensate for the weaknesses of others.

V. Conclusions

In this study, we investigated the limits of predictability for indoor human mobility. We examined multiple methods for modeling this predictability and provided evidence to support the models based on sensor messages collected from 117 smart homes. In single-resident smart homes, we found that the upper bound of the prediction accuracy for the next sensor message averages 83%. With the presence of multiple people in the smart home, the predictability lowers on average by 11%. If pets are present in smart homes, the predictability decreases by approximately 3% for both single-resident and multi-resident settings.

Although deriving a predictive model for a particular resident is beyond the scope of this paper, the above results provide an expectation of mobility prediction performance in real-world deployments. For applications that rely on the prediction of resident movement trajectories, developers can make more educated design decisions to achieve an improved user experience based on the statistical limits of mobility prediction performance.

Moreover, we can use the difference between the theoretical limits of predictability and the prediction accuracy of a particular mobility model as a quantitative absolute measurement of the performance of the mobility model. In the experiments, we assessed the performance of Markov chain-based mobility model in both single-resident and multi-resident settings. In terms of predictability, we found that the performance of Markov chain models is approximately 6% lower than the theoretical upper bound in single resident settings and 11% lower in multi-resident settings. The results indicate that more representative models could be developed for indoor human mobility, especially in multi-resident homes. To construct a mobility model that achieves a performance above the theoretical limit of predictability, information in addition to the residents' past trajectory, such as the time of day, the day of the week, and the length of resident stay at specific locations, will be needed.

A limitation of the current analysis is the dependence of the mobility states on the characteristics of the smart home sensors. A future analysis may consider a mobility metric that normalizes environment model based on factors such as home size, number of sensors, and sensor resolution, to lessen this dependency. Alternatively, sensor resolution can be integrated into future models, to aid in determining the number of sensors that needed to achieve the desired trade-off between predictability and sensor resolution.

The proposed models assume that human mobility is a stationary ergodic process. In many settings, this assumption will be violated. Mobility patterns may change due to seasonal impacts and changes in daily behavior patterns, among other contributing factors. While methods such as change point detection [6] may be used to identify such changes and initiate a new model, a future analysis may propose a model that does not rely on this assumption. Finally, future work may extend predictability to consider a larger horizon than one time point into the future. These improved capability will support more anticipatory support of human mobility for applications such as home automation, security monitoring, and activity prompting [24].

References

[1] L. Pappalardo and F. Simini, "Data-driven generation of spatio-temporal routines in human mobility," Data Mining and Knowledge Discovery, vol. 32, pp. 787–829, 2017.

[2] N. E. Klepeis, W. C. Nelson, W. R. Ott, J. P. Robinson, A. M. Tsang, P. Switzer, J. V. Behar, S. C. Hern, and W. H. Engelmann, "The national human activity pattern survey (NHAPS): a resource for assessing exposure to environmental pollutants," J. Exposure Sci. and Environmental Epidemiology, vol. 11, no. 3, p. 231, 2001.

[3] J. A. Leech, K. Wilby, E. McMullen, and K. Laporte, "The Canadian human activity pattern survey: Report of methods and population surveyed," Chronic Diseases in Canada, vol. 17, no. 3-4, pp. 118–123, 1996.

[4] C. J. Matz, D. M. Stieb, K. Davis, M. Egyed, A. Rose, B. Chou, and O. Brion, "Effects of age, season, gender and urban-rural status on time-activity: Canadian human activity pattern survey 2," International Journal of Environmental Research and Public Health, vol. 11, no. 2, pp. 2108–2124, 2014.

[5] A. R. Jimenez, F. Seco, P. Peltola, and M. Espinilla, "Location of persons using binary sensors and ble beacons for ambient assistive living," in International Conference on Indoor Positioning and Indoor Navigation, 2018.

[6] S. Aminikhanghahi, T. Wang, and D. J. Cook, "Real-time change point detection with application to smart home time series data," IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 5, pp. 1010–1023, 2019.

[7] D. Cook, A. Crandall, B. Thomas, and N. Krishnan, "Casas: A smart home in a box," IEEE Computer, vol. 46, pp. 26–33, 2013.

[8] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," Science, vol. 327, no. 5968, pp. 1018–1021, 2010.

[9] M. Lin, W.-J. Hsu, and Z. Q. Lee, "Predictability of individuals' mobility with high-resolution positioning data," in ACM Conf. Ubiquitous Computing, 2012, pp. 381–390.

[10] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, "Human mobility: Models and applications," Physics Reports, 2018.

[11] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," Scientific reports, vol. 3, p. 2923, 2013.
[12] G. Smith, R. Wieser, J. Goulding, and D. Barrack, "A refined limit on the predictability of human mobility," in IEEE International Conference on Pervasive Computing and Communication, 2014, pp. 88–94.
[13] A. Cuttone, S. Lehmann, and M. Gonzalez, "Understanding predictability and exploration in human mobility," EPJ Data Science, vol. 7, 2018.
[14] P. Baumann, C. Koehler, A. Dey, and S. Santini, "Selecting individual and population models for predicting human mobility," IEEE Transactions on Mobile Computing, vol. 17, pp. 2408–2422, 2018.
[15] P. Xu, L. Yin, Z. Yue, and T. Zhou, "Onpredictability of time series," Physica A: Statistical Mechanics and its Applications, vol. 523, pp. 345–351, 2019.
[16] S. Robben, G. Englebienne, and B. Krose, "Delta features from ambient sensor data are good predictors of change in functional health," IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 4, pp. 996–993, 2017.
[17] A. K. Sikder, L. Babun, and A. S. Ulugac, "Aegis+: A context-aware platform-independent security framework for smart home systems," Digital Treats: Research and Practice, vol. 2, no. 1, 2021.
[18] J. Dahmen, B. Minor, D. Cook, T. Vo, and M. Schmitter-Edgecombe, "Design of a smart home-driven digital memory notebook to support self-management of activities for older adults," Gerontechnology, vol. 17, no. 2, pp. 113–123, 2018.
[19] M. Khan, M. M. Saad, M. A. Tariq, J. Seo, and D. Kim, "Human activity prediction-aware sensor cycling in smart home networks," 2020.
[20] S. Wu, J. B. Rendall, M. J. Smith, S. Zhu, J. Xu, H. Wang, Q. Yang, and P. Qin, "Survey on prediction algorithms in smart homes," IEEE Internet of Things Journal, vol. 4, no. 3, pp. 636–644, 2017.
[21] K. Gopalratnam and D. J. Cook, "Online sequential prediction via incremental parsing: The active lezi algorithm," IEEE Intelligent Systems, vol. 22, no. 1, pp. 52–58, 2007.
[22] M. R. Alam, M. B. I. Reaz, and M. M. Ali, "SPEED: An inhabitant activity prediction algorithm for smart homes," IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, vol. 42, no. 4, pp. 985–990, 2012.
[23] K. Jayarajah and A. Misra, "Predicting episodes of non-conformant mobility in indoor environments," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2, 2018.
[24] B. Minor, J. Doppa, and D. Cook, "Learning activity predictors from sensor data: Algorithms, evaluation, and applications," IEEE Transactions on Knowledge and Data Engineering, vol. 29, pp. 2744–2757, 2017.
[25] V. Ghasemi and A. A. Pouyan, "Modeling users' data traces in multi-resident ambient assisted living environments," Intl. J. Computational Intell. Syst., vol. 10, no. 1, pp. 1289–1297, 2017.
[26] K. Doty, S. Roy, and T. R. Fischer, "Explicit state-estimation error calculations for flag hidden markov models," vol. 64, no. 17, pp. 4444–4454, 2016.
[27] B. Lin and D. J. Cook, "Using continuous sensor data to formalize a model of in-home activity patterns," Journal of Ambient Intelligence and Smart Environments, vol. 12, no. 3, pp. 183–201, 2020.
[28] R. Begleiter, R. El-Yaniv, and G. Yona, "On prediction using variable order markov models," Journal of Artificial Intelligence Research, vol. 22, no. 1, pp. 385–421, 2004.
[29] C. Dimitrakakis, "Bayesian variable order markov models," in International Conference on Artificial Intelligence and Statistics, 2010, pp. 161–168.
[30] "Algorithms for variable length markov chain modeling," Bioinformatics, vol. 20, no. 5, pp. 788–789, 2004.
[31] N. Roy, A. Roy, and S. K. Das, "Context-aware resource management in multi-inhabitant smart homes: a NASH H-learning based approach," in IEEE Intl. Conf. Pervasive Computing and Commun., 2006, pp. 148–158.
[32] T. M. Cover and J. A. Thomas, Elements of information theory. John Wiley & Sons, 2012.
[33] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with applications to english text," IEEE Transactions on Information Theory, vol. 44, no. 3, pp. 1319–1327, 1998.
[34] R. Takahira, K. Tanaka-Ishii, and Ł. Debowski, "Upper bound of entropy rate revisited—a new extrapolation of compressed large-scale corpora–," in Workshop Computational Linguistics for Linguistic Complexity, 2016, pp. 213–221.
[35] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," IEEE Transactions on Information Theory, vol. 23, pp. 337–343, 1977.
[36] J. Clear and I. Witten, "Data compression using adaptive coding and partial string matching," IEEE Transactions on Communications, vol. 32, pp. 396–402, 1984.
[37] V. Q. Vu, B. Yu, and R. E. Kass, "Coverage-adjusted entropy estimation," Statistical Medicine, vol. 26, no. 21, pp. 4039–4060, 2007.
[38] S. Aminikhanghahi and D. J. Cook, "Enhancing activity recognition using CPD-based activity segmentation," Pervasive and Mobile Computing, 2019.
[39] T. Wang and D. J. Cook, "Multi-person activity recognition in continuously monitored smart homes," IEEE Transactions on Emerging Topics in Computing, 2021.
[40] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in International Conference on Neural Information Processing Systems, 2013, pp. 3111–3119.
[41] H. Jun, S. Pan, M. K. Sinha, H. Y. Noh, P. Zhang, and P. Tague, "Smart home occupant identification via sensor fusion across on-object devices," ACM Transactions on Sensor Networks, vol. 14, p. Article No. 23, 2018.

Tinghui Wang received the B.E. degree in Microelectronics from Shanghai Jiaotong University, Shanghai, China and the Ph.D. degree in electrical and computer engineering from Washington State University. He is currently at Amazon. His research involves activity recognition, multi-target tracking in smart home environments with binary sensors.

Diane J. Cook is a Huie-Rogers Chair Professor at Washington State University. She received her Ph.D. in computer science from the University of Illinois and her research interests include machine learning, pervasive computing, and designed of automated strategies for health monitoring and intervention. She is an IEEE Fellow and a Fellow of the National Academy of Inventors.

Thomas R. Fischer received the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Massachusetts and the Sc.B. degree magna cum laude from Brown University. From 1989 until 2019 he was a Professor in the School of Electrical Engineering and Computer Science at Washington State University. He is currently Professor Emeritus. His research interests include data compression, image and video coding, digital communications, and digital signal processing.