4-2018

Predicting episodes of non-conformant mobility in indoor environments

Kasthuri JAYARAJAH
Singapore Management University, kasthurij.2014@phdis.smu.edu.sg

Archan MISRA
Singapore Management University, archanm@smu.edu.sg

Predicting Episodes of Non-Conformant Mobility in Indoor Environments

KASTHURI JAYARAJAH, Singapore Management University
ARCHAN MISRA, Singapore Management University

Traditional mobility prediction literature focuses primarily on improved methods to extract latent patterns from individual-specific movement data. When such predictions are incorrect, we ascribe it to 'random' or 'unpredictable' changes in a user's movement behavior. Our hypothesis, however, is that such apparently-random deviations from daily movement patterns can, in fact, *often be anticipated*. In particular, we develop a methodology for predicting Likelihood of Future Non-Conformance (LFNC), based on two central hypotheses: (a) the likelihood of future deviations in movement behavior is positively correlated to the intensity of such trajectory deviations observed in the user's recent past, and (b) the likelihood of such future deviations increases if the user's strong-ties have also recently exhibited such non-conformant movement behavior. We use extensive longitudinal indoor location data (spanning 4+ months) from an urban university campus to validate these hypotheses, and then show that these features can be used to build an accurate non-conformance predictor: it can predict non-conformant mobility behavior two hours in advance with an AUC ≥ 0.85, significantly outperforming the baseline. We also show that this prediction methodology holds for a representative outdoor public-transport based mobility dataset. Finally, we use a real-world mobile crowdsourcing application to show the practical impact of such non-conformance: failure to identify such likely anomalous movement behavior causes workers to suffer a noticeable drop in task completion rates and reduces the spatial spread of successfully completed tasks.

CCS Concepts: • **Information systems** → *Spatial-temporal systems*; *Data mining*; • **Human-centered computing** → *Empirical studies in ubiquitous and mobile computing*;

Additional Key Words and Phrases: Urban computing, spatio-temporal patterns, predictive modeling, location-based services

1 INTRODUCTION

The ability to predict an individual's future location (or indirectly, her movement behavior) is a key enabler of many mobile computing applications and services. In the past decade, there has been an explosive growth in the availability of large-scale mobility datasets (e.g., [6, 9, 18, 20, 21, 28, 38, 39]), obtained via technologies such as GPS (e.g., [9, 28, 38]), cellular records (e.g., [6, 18]) and WiFi logs (e.g., [20, 21, 39]), and capturing both campus and city-level movement. Researchers have used such datasets to empirically establish some of the scientific underpinnings of human mobility, including the predictability levels (or routine nature) of daily movement and the strong correlation between physical movement and social tie strengths. Two of the most typical prediction tasks investigated by researchers [4]) include NPP (*Next Place Prediction*–i.e., where will the user move next?)

Authors' addresses: Kasthuri Jayarajah, Singapore Management University, kasthurij.2014@smu.edu.sg; Archan Misra, Singapore Management University, archanm@smu.edu.sg.

Published in Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,
Volume 2 Issue 4, December 2018, Article No. 172.
https://doi.org/10.1145/3287050

and RT (*Residency Time*–i.e., how long will the user stay at the current location?). Such work fundamentally looks to uncover a variety of latent mobility patterns, and can enable a variety of predictive applications, such as anticipatory temperature control of a home [31] or proactive delivery of relevant alerts (by digital assistants such as Amazon's Alexa™ or Google's Google Assistant).

In this work, we tackle a distinct question: *Likelihood of Future Non-Conformance* (**LFNC**),–i.e., the odds that a user will not visit a routine place that she regularly visits. It is worth reinforcing, at the outset, the *distinctness* of our research question. Mobility prediction is principally about uncovering the underlying *routines* or patterns of a user, with the prediction accuracy being bounded by the inherent randomness (or predictability) of a user's movement behavior. Moreover, prediction algorithms focus on metrics such as minimizing average location error (i.e., the distance between the actual & predicted location coordinates). Instead, we embrace the fact that even the most predictable or routine user will, occasionally, diverge from such common mobility behavior, in an *apparently "random" or unpredictable* manner. For example, in a campus setting, a research group that has a regularly scheduled meeting on Thursday afternoons will "suddenly" skip a meeting. We thus focus principally on the question: *How can we enhance the confidence of declaring, sufficiently in advance, that a user will not be visiting a particular location, based purely on historical traces of location data?* Unlike mobility prediction, *LFNC* is measured by a more binary outcome variable: will the user actually be at the most-likely predicted location or not?

Our research focuses principally on movement behavior in workplace environments (e.g., a university or office campus), and arises from the observation that user mobility in a workplace is largely a manifestation of underlying, often-routine activities. In particular, significant research [35, 36] has been conducted on inferring or understanding workplace behavior via routine, scheduled activities and calendar events (e.g., group meetings, research discussions and lunches). Our investigations are motivated by the realization that the ability to predict such (likely rare) cases of non-conformance may become increasingly important in an age of anticipatory services, where predicting the wrong context might lead to greater negative consequences (e.g., unwanted or misleading notifications by a virtual assistant) than simply declaring "I'm unable to predict".

We develop a methodology to demonstrate that such "random deviations" can, in fact, be predicted, with a surprisingly high degree of accuracy. More importantly, we also explore the *lookahead capability* of such non-conformance prediction: i.e., we investigate the question "How far in advance can one reliably predict that a user (or users) will not be at a routine location?" Driven by past results that establish the strong social influence on mobility patterns, we shall investigate these questions by considering the impact of peer movement behavior on the predictability of such non-conformance. In fact, our research is driven by the following two hypotheses:

- H1–*Temporal Correlation of Non-Conformance:* A user's propensity of deviating from a future routine location/activity pattern is correlated to the anomalousness of her current movement–if a user has been exhibiting anomalous movement patterns in the recent past, she is much more likely to deviate from her routine location/activity pattern in the future;
- H2–*Homophily of Anomalies:* In workplace environments, where users indulge in significant collaborative activities, anomalous movement behavior is often not isolated but *shared*: if a user's "friends" have been exhibiting non-routine movement as well, there is a significant increase in the likelihood that she will deviate from her future routine movement pattern.

Key Contributions: We make the following major contributions:

- *Quantify Movement Predictability in Workplace/Campus Environments:* We use our primary campus-based location dataset (with its intrinsic tracking error of $\approx 6-8$ meters) to establish that the upper bound on the predictability of user movement in campus environments is comparable to prior results established for city-scale movement patterns. In particular, the median indoor predictability was 87% at the section-level granularity (compared to 91% in the outdoor case), with low variability across users (i.e., 99% of users

exhibited predictability greater than 80%). By additionally developing the mechanism to *deduce* social ties purely from such location data, we also establish the strong correlation between movement behavior and social ties in indoor workplaces, corroborating prior results on outdoor mobility [11].

- *Establish Lookahead & Predictability of LFNC:* We determine the lookahead capability (how far in advance can we deduce that a user will deviate from a regular movement pattern) by considering both (i) the lookahead distance and (ii) the length of the current movement sequence. We demonstrate a sufficiently strong lookahead capability of $\approx$ 2 hour on a typical workday, with an $AUC \geq 0.85$, using possible anomalous behavior observed over the last 15 minutes. (This represents an improvement of 15+% over a baseline classifier that simply looks at the probability of the most likely predicted location.) We also show that the inclusion of anomalies occurring too far in the past (e.g., over the past 4 hours) leads to a substantial loss of prediction accuracy.
- *Demonstrate & Harness the Collective Nature of LFNC:* We empirically establish that the movement patterns of strong social ties are correlated, not just during regular movement, but *also during episodes of anomalous movement*. We harness this correlation in an anomaly predictor, that utilizes an easy-to-implement supervised machine learning technique and show that we can significantly improve the AUC of LFNC prediction (over an individual-centric predictor), for a lookahead time of 3 hours, by up to 15% (with a final $AUC \approx 0.90$), with comparable improvements in precision and recall.
- *Robustness and Evolution of LFNC Prediction:* Through careful numerical studies, we show that our proposed LFNC predictor is applicable across a wide variety of scenarios. We show that our classification accuracy is robust, maintaining high AUC even when the classifier is trained on data that has no temporal overlap with the test instances. The predictor also has a fairly short *cold-start* period: it is able to infer imminent anomalous mobility behavior after observing user movement patterns on campus for only 2-3 weeks. Moreover, our method of predicting LFNC, using a combination of anomalous past movement and anomalous behavior of social ties, is applicable even for outdoor city-scale mobility. Using trip data on a public transit system, we show that we can predict anomalous commuting behavior with AUC ≥ 0.85, an improvement of over 30% over an anomaly-oblivious baseline.
- *Improved Efficiency for Location-Predictive Applications:* Using a real-world trace-driven study, we demonstrate that our LFNC prediction can provide compelling practical benefits. We apply LFNC to real-world traces of a campus mobile crowd-sourcing application. We show that we can achieve a high AUC of 0.90 in the ability to predict deviations of individual crowd-workers from their regular *staypoints* with 1-hour look-ahead windows – we infer additional insights that suggest that workers with *anomalous* or deviant trajectories perform tasks with less spatial diversity (a 12.5% decrease in entropy over the task locations that they chose) and also exhibit overall lower task completion rates.

Overall, we believe that we are the first to (a) scientifically quantify the predictability of future deviations from routine movement behavior, and then (b) demonstrate the practical benefit of such anomaly prediction for common, real-world applications.

2 PRELIMINARIES

In this section, we describe the definitions and notations used throughout the draft, and the indoor location data sets used in this work.

2.1 Approach at a Glance

Our primary goal in this work is in validating our two central hypotheses: (1) the deviations a user exhibits from his/her routine behavior in the past can be useful predicting future non-conformance, and (2) that the concurrently, deviating behavior of his/her social network can improve that predictability. To this end,

- **Step 1:** We first study longitudinal indoor mobility data, from an urban campus, in order to shed light into two important prerequisites to operationalize our hypotheses (Section 3). In particular, we investigate the (1) theoretical limits to predictability in indoor settings, and the (2) evolution and stability of physical social ties in the campus setting.
- **Step 2:** Next, we propose a Likelihood of Future Non-Conformance (LFNC) prediction pipeline that relies on a supervised learning classifier (Section 4).
- **Step 3:** We then evaluate the proposed pipeline, extensively; we study the trade-off between the look-ahead distance (how far into the future), impact of social ties and the LFNC performance. We conduct several experiments to validate the robustness of the prediction pipeline (Section 5).
- **Step 4:** We study the practical usefulness of LFNC predictions using data from a route-aware mobile crowd-tasking system operational in an urban campus.
- **Step 5:** Finally, we study the extensibility of the proposed pipeline to the outdoor setting using city-scale, public transit trip data.

2.2 Definitions and Notations

We consider a trajectory $x_(u,d) := \{loc_1, loc_2, ..., loc_T\}$ whose elements are a sequence of staypoints loc_t, a user u visited during a day d, and $t \in [1, T]$. $x_(u,d)$ is a $T-$length vector each element representing equally sized, time bins over the day – for instance, for a trajectory considered at the hourly granularity, $|x_(u,d)| = 24$.

Next Place Prediction: Given a collection of $x_(u,d) \in X_{train}$ where X_{train} is the mobility training period (see Figure 1), and the trajectory of the same user u till time t, on a different day d_{test} outside the training period (i.e., $x_{(u,d_{test},1:t)} \in X_{test}$), the most likely next place at time $t+1$ denoted by $np_{u,d,t+1}$ can be predicted using a next place prediction algorithm. Baumann et al. [4] provide a survey of 18 such prediction algorithms and report on their performance.

Non-Conformance: We declare non-conformance at such time $t+1$, when $loc(u, d_{test}, t+1) \neq np_{u,d_{test},t+1}$. Further, we define a look-ahead distance, K.

Future Non-Conformance: Then, non-conformance at a future time $t+K$ is defined by Equation 1.

$$Nonconformance(u, d_{test}, t, K) = \begin{cases} 1, & \text{if } loc_{(u,d_{test},t+K)} \neq np_{(u,d_{test},t+K)} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Trajectory Deviation: We define a current sequence, $x_{(u,d_{test},t-h:t)}$, of length h, as at time t. A user's *deviation*, $d_{(u,d_{test},t,h)}$, from his/her routine trajectory, is then given by Eq. 2. Here, $dist(.)$ measures the time series distance between the two partial trajectories. Note that $\hat{x}_{(u,d_{test},t-h:t)} := np_{(u,d_{test},t-h)} \| ... \| np_{(u,d_{test},t)}$.

$$d_{(u,d_{test},t,h)} = dist(x_{u,d_{test},t-h:t}, \hat{x}_{u,d_{test},t-h:t}) \tag{2}$$

Social Ties: Separately, we construct a physical social network of users based on their movement trajectories in X_{Train}, adopting the definitions from Jayarajah et. al [20]. For each user u, we consider his/her k ties with whom they share the highest cumulative tie strength as u's *ego network*, $U_k \subset U$. Further, we denote the deviation the user u's top-k ties concurrently by the set $D_{u,d_{test},t,h,k}$ whose elements are $d_{u_k,d_{test},t,h}$ where $u_k \in U_k$.

Likelihood of Future Non-Conformance: Then, we define LFNC as the probability of occurrence of a future non-conformance ($Likelihood(loc_{(u,d_{test},t+K)} \neq np_{(u,d_{test},t+K)})$), as at time t on d_{test}, given $d_{u,d_{test},t,h}$, i.e., the user's deviation from norm thus far (limited by h), and $D_{u,d_{test},t,h,k}$. As we describe later in Section 4, we operationalize LFNC as a classification task whose outcome variable is a probability that is equal to $Likelihood(loc_{(u,d_{test},t+K)} \neq np_{(u,d_{test},t+K)})$.

Table 1. Notations used throughout the text.

Notations	Meaning
U	Set of all users
K	Look-ahead distance
k	Number of social ties
U_k	User u's ego network of top-k users
h	Current sequence length
T	Trajectory length
$dow(d)$	Day of the week of d
$x_{(u,d)}$	Trajectory of user u on day d
$loc_{(u,d,t)}$	Location of u at time $t \in [1, T]$) on day d
$x_{(u,d,t-h:t)}$	Current sequence over which deviation is computed; partial trajectory of (u, d) during time $[t - h, t]$
$\hat{x}_{(u,dow,t-h:t)}$	Expected (or, routine) current sequence of user u for the same day of the week as d
$d_{u,d,t,h}$	Deviation during the current sequence, $dist(x_{(u,d,t-h:t)}, \hat{x}_{(u,dow,t-h:t)})$
$np_{(u,d,t+1)}$	Most likely *next place* of user u on day d given trajectory till t

2.3 Dataset Description

Our dataset includes the location traces of individuals residing or visiting our university campus. The university is located in the downtown of a major Asian city, and comprises approximately 10,000 students (undergraduate and graduate) and 1,500 staff. The university has no on-campus residential facilities; hence, all campus inhabitants commute to/from their residence. The university comprises 7 academic buildings, 1 administration building and an underground 'concourse' that acts as a publicly-accessible connector between the academic buildings.

The indoor location data is generated using a WiFi fingerprinting-based indoor location system, which has been operational on the campus for over 3 years, and which covers all the publicly-accessible parts of the 7 academic buildings and the underground concourse. The location system uses fingerprint measurements taken at *landmarks*: with modest exceptions (to accommodate irregular building layouts), landmarks are spaced 3 meters apart.The WiFi-based system utilizes the RSSI readings, from each WiFi-enabled device resident on campus, to compute the device's *medium-grained* indoor location, achieving a median accuracy of 6-8 meters (2-3 landmarks). Because of this medium-grained location tracking, we often express the location coordinates of each user at *section-level* granularity: a section typically corresponds to a collection of landmarks, and represents a logical partitioning of a building floor (e.g., a classroom, a group-study (meeting room), a food outlet, etc.).

New location estimates are generated once approx. 5-10 seconds. To focus primarily on the personal mobile devices of regular campus residents, we filter out (i) devices such as laptops and desktops (that exhibit only sporadic, intermittent mobility) and (ii) devices that belong to transient campus visitors (we require the device MAC address to be seen on campus at least for over 10 minutes over a day). Note that, due to the growing trend for devices to perform MAC address randomization when in a disconnected state, we effectively filter out those devices that do not connect to our WiFi network.

We use data from two time periods: (1) *Dataset A* in which the set of users on campus whose periodic outdoor mobility information was also available (which we utilize in Section 3 for comparison in predictability), and (2) *Dataset B* consisting of users who participated in a campus-wide crowd-tasking pilot (whose details we utilize in Section 6). Table 2 summarizes key details. As illustrated in Figure 1, we split *Dataset B* into disjoint mobility model training (X_{Train}) and test (X_{Test}). As noted previously in Section 2.2, Next Place Predictions are made over X_{Test} which then serves as the dataset for LFNC learning and classification.

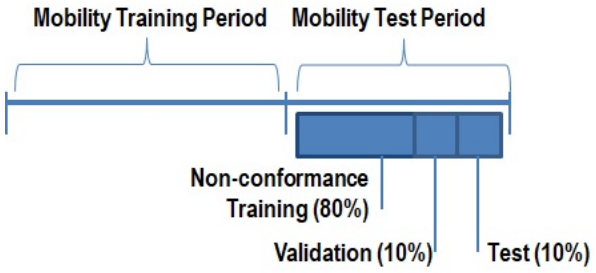


Fig. 1. Illustration of the dataset segregation for different train/test purposes.

Table 2. Overview of Data Sets used in this Work.

	Observation Period	Users ($\|U\|$)	Section
Dataset A	Aug–Dec, 2014	36	Section 3
Dataset B	Feb–Mar, 2017	806	Section 5, 6

3 CAMPUS-SCALE MOBILITY: EMPIRICAL INSIGHTS

Our broad goal is to predict deviations/anomalies from a user's regular movement pattern, using the collective movement pattern of an individual and her 'social ties' to improve the prediction accuracy. In addition, we would like to understand the lookahead time of such predictions, and how ongoing/past anomalous movement influences the prediction of future anomalies. However, for this approach to be successful, there are a few fundamental questions & challenges that we need to resolve first:

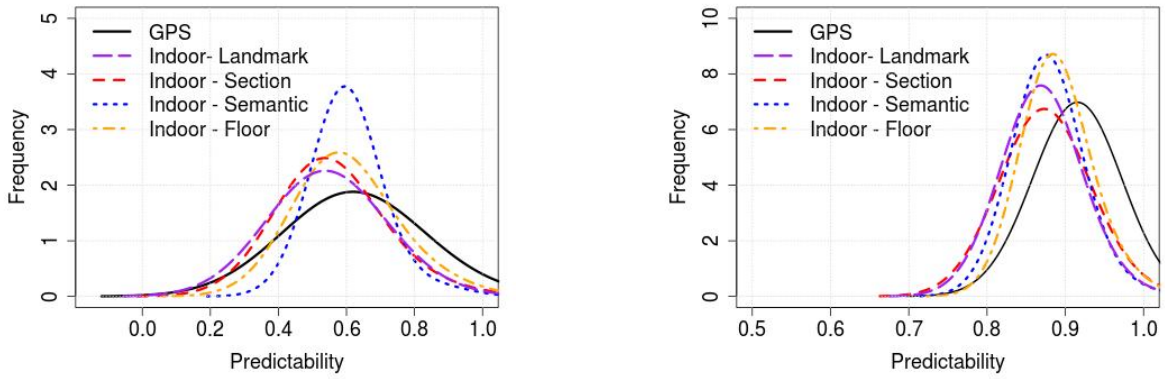
- What are the fundamental limits on predictability for indoor movement behavior in a workplace/campus setting? How does it differ from prior results on outdoor human mobility? In particular, for indoor environments, where the location trace itself has moderate error, the unpredictability is driven by both the random properties of human movement, and the noise introduced by the location traces.
- How do we reconstruct or infer social ties solely from the collective location traces of individuals? (Recall that our goal is to utilize social ties built unobtrusively from the physical location traces, unlike past work that constructs such ties from other observations (e.g., online social network traces [9] or call records [29]). And, how do we verify that our inferred social ties are meaningful and stable?

3.1 Predictability Indoors

Fundamentally speaking, an individual's trajectory can be seen as a random sequence of symbols (with each landmark being a distinct symbol). The degree of randomness in such a movement pattern can be computed based on the notion of entropy rate of this random sequence. We borrow concepts from Song et al.[33] in defining the theoretical upper bound on predictability of outdoor mobility using cell associations and explore predictability *indoors*.

In Figure 2, we plot the distribution of predictability [21, 33] across users based on the (a) uncorrelated entropy where only the probability of a user turning up at a location is known (Figure 2a), and (b) the true entropy where the full history of a user's spatial and temporal patterns are known (Figure 2b), for users in *Dataset A*. We observe at varying spatial granularity of localization – in particular, at the (1) landmark level (most fine-grained, every 3 meters), (2) section level (6-8 meters), and (3) floor level, and compare against the predictability outdoors, observed via the continuous reporting of GPS coordinates (rounded to the third decimal which results in a granularity of ≈ 100 meters), for the same set of users.

In general, we observe that the maximum predictability is comparable to the outdoor case – for instance, the median predictability for GPS is 91% while the same is 87% at the section level. Noticeably, unlike what was reported in Jensen et al. [21] where predictability was looked at at the raw WLAN association levels indoors, we observe that at coarser spatial granularities (e.g., section level), the variability across users also reduces. For instance, while the previous work reported a non-negligible percentage of users showing maximum predictability



(a) Uncorrelated Trajectory Sequences

(b) Infinite Length Sequences

Fig. 2. Theoretical Maximum Predictability at Varying Spatial Granularity for (a) Temporally Uncorrelated, and (b) Correlated Trajectory Sequences

in the range of 0.2 to 0.9, at the floor level, we observe that 99% of the users are bound to within a narrow range of 0.80 to 0.98 (i.e., within 3 standard deviations away from the mean).

3.2 Social Ties and their Impact on Mobility

We first look at the formation and evolution of social ties in our campus setting in order to understand it's practical usefulness for accounting friendship information as additional features that reveal insights into a user's mobility. We note that the 36 users in *Dataset A* are all freshman, and we observe their mobility since their first week on campus till the end of the term.

As described previously in Section 2.2, we use the technique described in Jayarajah et al. [20] to infer the intensity or strength of tie between any two users. In Figure 3, we plot several metrics related to the evolving *friendship network* amongst the users, as the term progresses (represented by $x-$ axis). Between consecutive weeks, we consider the sub-network consisting of only the top-1 ties (i.e., the network consisting of each user and his/her closest tie, till current week), and observe that the Jaccard similarity [27] of the set of edges, steadily reaches its maximum at week 5 after which it plateau. Similarly, we see that the diameter (represented by the blue dotted line) of the complete network, undergoes a stark drop till week 3 after which it stabilizes. This shows that the network consisting of the closest ties stabilizes after the initial few weeks, and demonstrates that the passively captured social ties can in fact be reliably used as additional information in exploring the predictability of user mobility.

Additionally, we studied the *correlation* of mobility of a user and his/her top-1 tie. We divided *Dataset A* as X_{Train}: Aug-Oct, 2014, and X_{Test}: Nov, 2014. For each day d, and time t in X_Test, we computed the *probability of being at* $loc(u, d, t)$, for each user u, based on the visitation frequency distribution over all possible locations learned during X_{Train}. We then compute the Pearson's correlation coefficient of the time series of *probabilities* of user u and his/her close tie, as well as the correlation between the same user and any randomly chosen user from the dataset. In Figure 4, we plot the CDF of the correlation values – we see that the similarity in being at likely locations (and not necessarily the same location) concurrently, is statistically significantly higher for top-1 pairs ($D = 0.395$, p-value = 0.005 on the Kolomogrov Smirnov test).

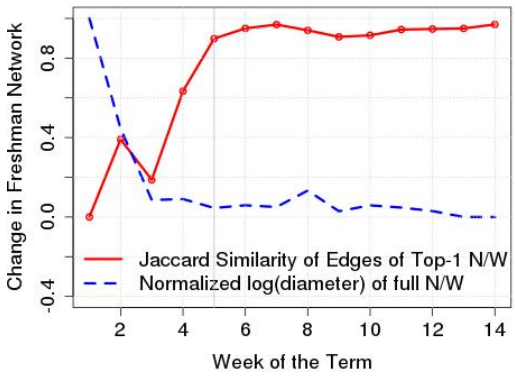


Fig. 3. Evolution of the Physical Friendship Network of Students, as the Term Progresses.

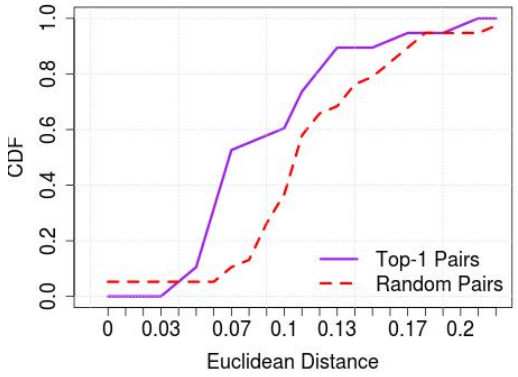


Fig. 4. CDF of correlation between pairs of trajectories belonging to top-1 ties vs. random pairs.

The correlation value is computed against the sequence of the probabilities of a user being at a location (and not the correlation in the location trajectory itself)

Key Insights: Our empirical analyses establishes 3 key results:

(1) User movement, in a mostly-indoor campus environment, has high predictability at the section-level and floor-level granularity, theoretically, and is reasonably consistent across users.
(2) The set of top-K ties (derived from an initial observation period of 5 weeks) remains remarkably stable over the remaining 9+ weeks of the term. Accordingly, we've established that it is possible to derive the set of 'strong ties' of an individual, unobtrusively, using a modest period of observational data.
(3) There exists significant correlation between the movement behavior/trajectory of close ties–i.e., "birds of a feather flock together". This finding corroborates similar insights previously presented for outdoor mobility [9, 11], and suggests that factoring in the movement behavior of close-ties should improve the location prediction accuracy for an individual.

4 LFNC PREDICTION PIPELINE

In this section, we describe the overall working of the non-conformance monitoring pipeline.

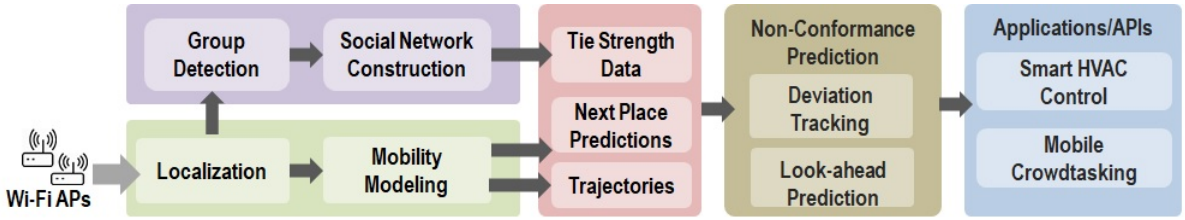


Fig. 5. Proposed Pipeline for LFNC Prediction.

4.1 Trajectory Extraction

As described previously in Section 2, the server-side indoor localization server receives reports on RSSI from individual APs as well as via the Real-Time Location Service (RTLS) running on the master controller allowing for frequent updates on the localization of individual mobile devices using an extension of the RADAR [3] algorithm. The *Localization* module then constructs trajectories per user, per day (i.e., $x_{(u,d)}$s) as defined in Section 2.2.

4.2 Mobility Modeling

This module implements three "next place" predictors: zeroR and Markov Chains of length 1 and 2 [15] (MC-1 and MC-2, respectively) taking into account the day of the week of day d (i.e., $dow(d)$) and time of the day t; in our implementation, we consider t at intervals of 15-minutes resulting in $T = 96$. Based on observations in X_{Train}, the constructed transition matrices for $MC-2$ and $MC-1$, and the visitation frequency matrix for $zeroR$ predictions allow for *next place predictions* in X_{Test}.

Similar to the implementation described in Kotz et al. [34], for each test sample, we roll back from $MC-2$ to $MC-1$ to $zeroR$, depending on whether the same context was *seen* during training. The matrices are stored as $key-value$ pairs, where the key is the concatenation of $< dow, t, loc_{t-1}, loc_t >$ for $MC-2$, $< dow, t, loc_t >$ for $MC-1$ and $< dow, t >$ for $zeroR$ predictions, respectively. During test time, for a given context at time $C_t :=< user, dow, t, loc_{t-1}, loc_t >$, if a MC-2 prediction is not possible (due to that case not seen during train time), the algorithms rolls back to $MC-1$; and if such a prediction, too, is not possible, the algorithm rolls back to $zeroR$. If $zeroR$ is also not possible, a prediction is not made.

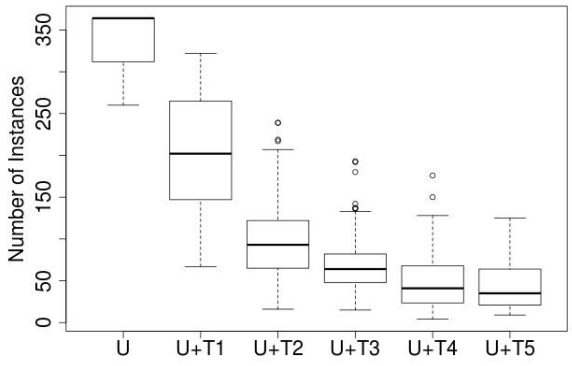


Fig. 6. Number of instances in the dataset with user alone vs. with friends simultaneously present on-campus, **per user**.

4.3 Tie Strength Extraction

We implement and use a state-of-the-art group detection system, GruMon [19], which has been demonstrated to accurately identify social groups using passive indoor location data. Based on longitudinal observations of such group episodes, as detailed in Section 2, we adopt the tie strength extraction method outlined by Jayarajah et al. [20] to measure the strength of tie between pairs of users in the dataset. In Section 5, we restrict our evaluations to the impact of a user's ego network consisting of only the top-k ties with k set to $\{1, 2, 3, 4, 5\}$. Although higher values of k is possible, it reduces the size of the dataset on which evaluations are possible as the number of cases

with top-k users are all present on-campus at the same time drops significantly, as k is increased. We show this in Figure 6.

4.4 Non-Conformance Prediction

Labelling Nonconformance: During the training phase, the *Look-ahead Prediction* module consumes pre-trained next place predictions for labeling whether the user will *conform* to, or be at the most likely place, at a time, $t + K$, i.e., K bins into the future.

As identified by the two central hypotheses of this work, we factor in: (1) the deviation, $d_{(u,d,t,h)}$ in trajectory a user has incurred during the *current sequence*, $x_{u,d,t-h:t}$, and (2) the deviation in trajectory the user's top-k friends have incurred, $D_{(u,d,t,h,k)}$.

Deviation Computation: We compute this deviation ($d_{(u,d,t,h)}$) as at time t as the distance between the time series of the actual trajectory a user has traversed during time $[t - h, t]$ and the sequence of *most likely locations* for the same period. To measure this distance, among the alternatives considered, Dynamic Time Warping (DTW) [26], Hamming Distance [17] and Edit Distance [30], we found the former to be the most appropriate.

To quantify the deviation of the user's friend network consisting of his/her top-k friends, we extract instances where all k friends are on-campus concurrently (and not necessarily co-located), and sum the individual deviations weighted by their strength of tie. In Section 5, we vary k to observe its impact on the performance of non-conformance prediction. As a direct consequence, with increasing k, the number of instances where the user and the top-k friends are all concurrently present on campus drops drastically.

Algorithm 1 outlines the steps taken to label nonconformance and compute the corresponding deviation measures.

Predicting Nonconformance: We train and build a Gradient Boosting Classifier whose independent variables are $< dow(d), t, d_{u,d,t,h}, D_{u,d,t,h,k} >$ and the binary outcome variable represents conformance (or, nonconformance), for different look-ahead distances of K.

Finally, the conformance predictions can be consumed by various applications including smart HVAC control and route-aware mobile crowdtasking (see Section6).

5 EVALUATION

In this section, we report our findings on the predictive ability of the individual factors based on the two central hypotheses that we consider in this work, in predicting future non-conformance of a user's mobility behavior. We first evaluate in Section 5.1 the influence of trajectory deviations a user and his/her ties undergo over a day, in predicting future non-conformance. Then, in Section 5.3, we explore the impact of time of the day and the types of places a user visits on such predictive performance. Further, we conduct a number of robustness checks in Section 5.4 – in particular, we answer the following additional questions:

(1) does the performance remain robust when evaluating cases only where the user *transitions* to a different staypoint?
(2) does a dynamically learned tie strength metric (over the course of the term) still useful?
(3) does the performance remain stable when train/test data are obtained from completely different days?

Prediction task: We represent the non-conformance prediction task as a binary classification task with the conformance label (1 – non-conformance and 0 – conformance) as the dependent variable and the features described in Section 4 as independent variables with a Gradient Boosted Model (GBM) for supervised classification. The choice of this classifier is motivated by the popularity of ensemble learning techniques and their demonstration in mobility prediction tasks in recent works [14, 32]. Later in Section 8, we compare the performance of other machine learning algorithms including Random Forest which is also an ensemble method.

Algorithm 1 LFNC Labeling and Deviation Computation

Input: $MC-2TransitionMatrix, MC-1TransitionMatrix, ZeroRMatrix, x_{(u,d)}, t, K, h, U_k$
Output: $d_{(u,d,t,h)}, label_{u,d,t,K}$
$maxK \leftarrow K - t$ ▷ The number of look-ahead windows possible after current time t
$predictedtrajectory \leftarrow trajectory_{1:t}$ ▷ Input only known trajectory till time t
$d_{(u,d,t,h)} \leftarrow NULL$ ▷ Initiate deviation vector
$label_{u,d,t,K} \leftarrow NULL$ ▷ Initiate conformance vector
for $K = 1$ `to` $maxK$ **do**
  $nextlocation_{u,d,t+K} \leftarrow getNextLocation(x_{(u,d,1:t+K-1)})$ ▷ Predicted *next location* at time $t + K$
  if $nextlocation_{u,d,t+K} == loc_{(u,d,t+K)}$ **then**
    $label_{u,d,t,K} \leftarrow 1$ ▷ Label conformance
  else
    $label_{u,d,t,K} \leftarrow 0$
  end if
  if $K == 1$ **then**
    $\hat{x}_{(u,d,t+K)} \leftarrow nextlocation_{u,d,t+K}$ ▷ Append expected next place
    $d_{(u,d,t,h)} \leftarrow dist(x_{(u,d,t-h:t)}, \hat{x}_{(u,d,t-h:t)})$ ▷ Compute deviation
  end if
end for

Experiment conditions: In all the experiments described in this section, we perform classification on a balanced set (with equal number of conformance, and non-conformance classes) derived from *Dataset B* (see Section 2.3); to do this, we first create a subset of all the samples from the smaller class and randomly sampled, equal sized samples from the other class, generating the balanced dataset. Unless explicitly stated, we take X_{Train} (01-02-2017 – 28-02-2017) over which the next place prediction models are trained, and X_{Test} (01-03-2017 – 14-03-2017) over which LFNC is trained and tested using a split of train (80%), validation (10%) and test (10%).

Parameter tuning and model selection: On the train set, we learn multiple GBMs assuming a Gaussian loss function and by varying the number of trees between 100 to 10,000 (in increments of 100). The shrinkage and interaction depth parameters are fixed to defaults (i.e., 0.01 and 4, respectively). We pick the best performing model as the one that has minimal error on the validation set. Finally, we evaluate the test set on this model to report findings.

Performance metrics: In all our analyses, we report the accuracy based on precision, recall and AUC, following their standard definitions. Precision and recall represent the average over both the positive and negative classes and we use 0.5 as the cut-off probability in declaring the binary outcome variable.

Implementation: The computations related to GBMs were performed using R's *gbm* package [1] and the ROCR [2] library for performance calculations.

5.1 Predicting Non-Conformance

We first evaluate the performance of predicting non-conformance under the three conditions: (1) $Model_{nodev}$, without any information of a user's deviation from his expected trajectory (i.e, with the day of the week, *dow*, and time bin, t, being the only input features), (2) $Model_{userdev}$, considering the deviation of the user in addition (i.e, $user_deviation$), and (3) $Model_{combidev}$, considering the deviation of both the user and his/her friends present on campus at that time (i.e., $d_{(u,d,t,h)}$ and $D_{(u,d,t,h,k)}$).

In Figure7, we plot the accuracy (measured as AUC, on the $y-$axis) for LFNC predictions with $Model_{userdev}$, for varying look-ahead distances K, and current sequence lengths, h, or in other words, the deviations from

Table 3. Prediction Results with no deviation, using only user's deviation, and the combination of user+friends' deviation.

	$Model_{nodev}$	$Model_{userdev}$	$Model_{combidev}$
N @ K=1		1096	
Precision @ K=1	0.665	1	1
Recall @ K=1	0.654	1	1
N @ K=4		1072	
Precision @ K=4	0.642	0.862	0.897
Recall @ K=4	0.642	0.858	0.896
N @ K=12		958	
Precision @ K=12	0.69	0.631	0.792
Recall @ K=12	0.66	0.625	0.792

how far back in the user's trajectory. We observe that observing trajectory deviations over short hs (e.g., 15 - 45 minutes), tend to result in better performance – for instance, for most values of K, the predictions corresponding to $h = 1$ and $h = 3$ (i.e., blue and red lines) exhibit performance improvement of over 10% consistently.

In Figure 8, we plot the accuracy with the AUC on the $y-$axis for varying look ahead times, K, on the $x-$axis, for the three cases where both the user and his/her top-5 ties were present on campus. Note that each increment in K implies the addition of 15 minutes into the future from current time. We observe that among the three, considering both the user's and friends' deviation thus far (represented by the solid blue line) offers the greatest performance – even with a look-ahead time of 3 hours (i.e., $K = 12$), the combination provides an $AUC =\approx 0.9$. We further note that considering the user's deviation alone performs similarly well until around $K = 4$ after which the drop off rate increases resulting in at least a 15% drop in accuracy in comparison to considering the friends' deviation in addition. We also note that the performance of not considering any of the deviation measures results in the poorest performance (relatively stable at $AUC =\approx 0.7$) – which means that the additional factors provide a performance improvement of ≈35%, 25% and 15% at look-ahead times $K = 1, 4, 12$, respectively. In Table 3, we tabulate the precision and recall values for the three cases.

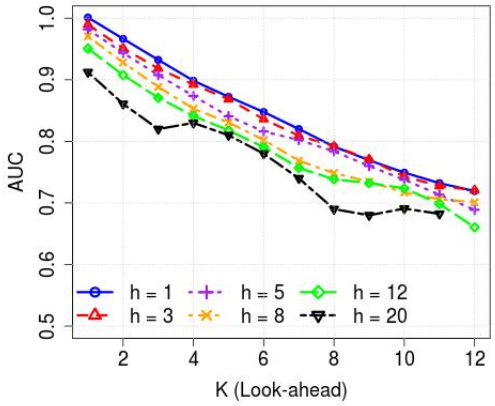


Fig. 7. Comparison of LFNC Look-Ahead Capability for different h.

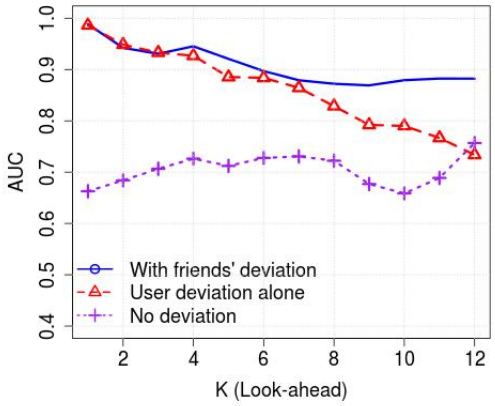


Fig. 8. Comparison of LFNC Look-Ahead Capability of different Models.

5.2 Impact of Social Ties

Previously, we noted the utility in considering the combined deviation of a user's social ties from their respective "expected" routines in predicting non-conformance, ahead of time. Here, we explore the impact of the "size", or

“extent" of the social ties considered on prediction performance. In Figure 9, we observe the performance of using user's deviation alone vs. user's top-k ties' combination of deviation, where we vary k from 1 to 5. We note here that, unlike in the previous case where we compare the performance against a subset of instances where the user and all top-5 ties were concurrently present on campus, here, we report the performance for the instances where top-k ties were concurrently present – in effect, this means that the number of samples trained on and evaluated against for the top-1 case is higher than that of the top-5 case (as we previously saw in Figure 6). We make the following remarks:

(1) Surprisingly, the consideration of a user's top-1 tie's deviation in addition to his/her own does *not* provide additional utility. However, we find that this is consistent with our prior observation (in Section 3) that the user and the closest ties' mobility behavior are too highly correlated that they do not provide additional information gain.
(2) We further note that with increasing size of the ego network, the performance improves – for $K \geq 3$, we see that the AUC is generally ≥ 0.80 for as advanced as 10 hours of look-ahead (i.e., $K = 40$) whereas the same for the user, or user and the closest tie combination stabilizes around 0.70.

Further, we investigated the performance in predicting non-conformance for times (i.e., $< dow, t >$ combination) during which the user has had differing levels of *regularity*, historically. The regularity is captured as the *zeroR* probability as in [33]. We plot the performance for subsets of user instances thresholded by varying values of the regularity, S (ranging from 0.3 to 0.7), in Figure 10a and Figure 10b, for $Model_{userdev}$ and $Model_{combidev}$ with k=5, respectively. We highlight that the performance improves in general (about 5%) over “more regular" periods (i..e., higher values of S), in either case.

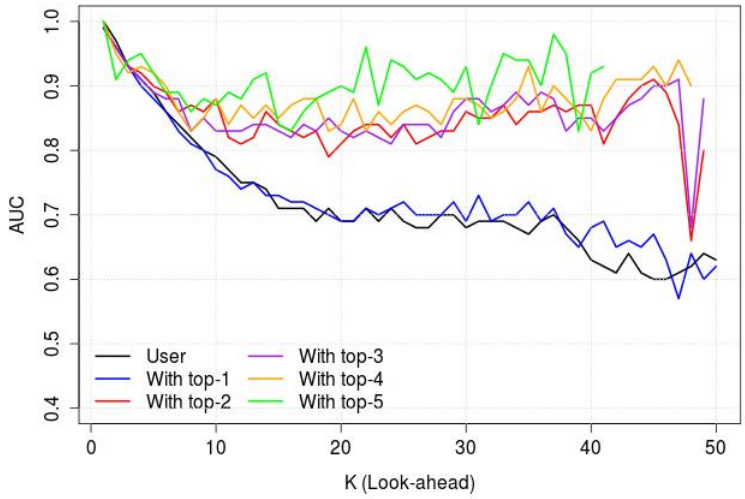


Fig. 9. Performance by varying size of the ego network of a user considered (and hence, the resulting combination deviation feature).

5.3 Performance by Time and Place

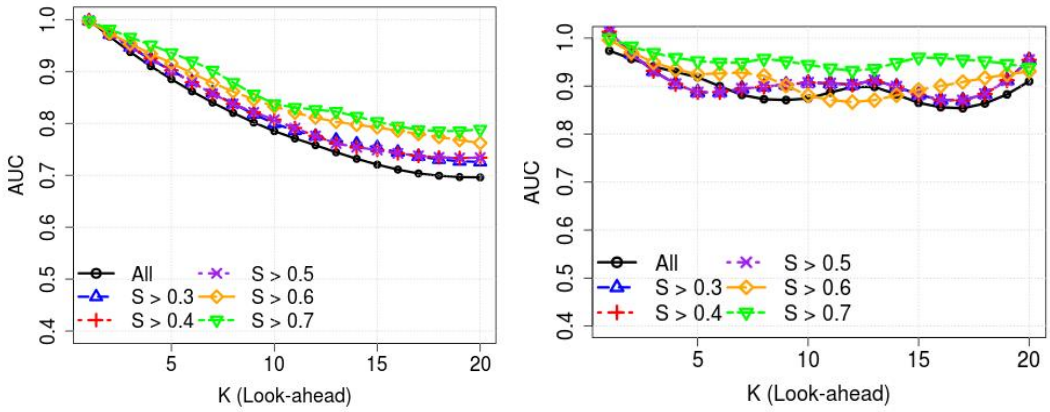
Thus far, we have reported the performance across all time windows and likely location types – here, we explore the role of time and place on performance, more closely.

In Figure 11a, we plot the performance of $Model_{nodev}$ and $Model_{combidev}$ stratified by the "time" – we consider Morning hours as between 8 AM to 12 Noon, Afternoon as 12 Noon to 5 PM and Evening as 5 PM to 9 PM. The "time" here represents the future time for which a prediction is made (i.e., current time t + look ahead time K). We observe the largest difference during morning hours where $Model_{combidev}$ performs at least 30% better in comparison, for up to look-ahead times of 3 hours ($K = 12$). We also note that the least difference in performance is observed for afternoon predictions – and that the gap keeps bridging with increasing K. For both models, we observe the steepest drop in performance with K for evening predictions where the performance of both drops as low as 0.6 for $K = 12$.

Further, in Figure 11b, we plot the performance stratified by two of the most common types of locations on campus – (a) seminar rooms (including any other scheduled teaching rooms such as class rooms), and (b) study areas (including group study rooms which are available to students as "booked" resource which is typically used for project discussions, and open study areas). Similar to the case above, the location here refers to the actual "future" location the user was at $t + K$. Between the two, the former represents a formal class of locations and the latter more a casual setting. As expected, we observe that the improvement in performance in considering the user's social ties' mobility behavior is evident for the more casual/social scenario – for e.g., a 40% improvement in prediction for $K = 1$ and which tapers down to roughly 25% with a look ahead time of 3 hours. For the most part, we observe that for the more formal setting, the performance of both models are comparable in that the impact of social ties' mobility has less impact on class attendance.

5.4 Robustness Checks

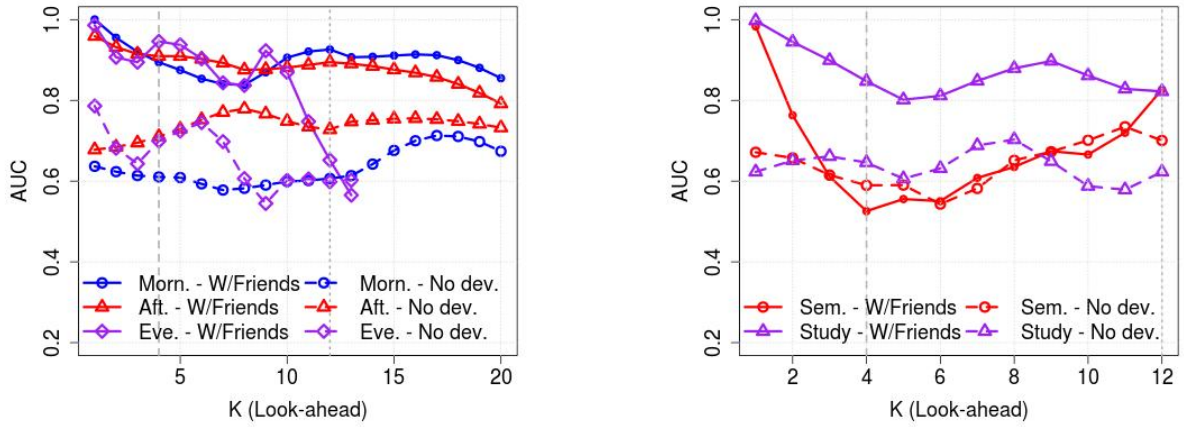
In this section, we report findings from a number of checks, in order to validate the robustness of non-conformance prediction under varying conditions.



(a) User's deviation alone $Model_{userdev}$

(b) User + Friends' deviation $Model_{combidev}$

Fig. 10. Performance by differing levels of regularity of the predicted time instance, historically, for (a) $Model_{userdev}$ and (b) $Model_{combidev}$ with $k = 5$.



(a) Accuracy by Time of Day

(b) Accuracy by Types of Places

Fig. 11. Difference in Performance by (a) the time of day and (b) the type of places a user is at.

5.4.1 Performance during Staypoint Transitions. Whilst previous work on outdoor mobility have demonstrated that the theoretical maximum predictability is achievable [33] in practice, there's a lack of evidence in the indoor setting. Kotz et al. [34] note that practically achievable accuracy of predicting next place is comparably less, and also that the prediction task is easy when the user is still (where the next place is the same as the current place) but suffers during *transitions*. To understand how this impacts future non-conformance predictions, we report on the performance for instances where *transitions* have occurred, in Figure 12. We distinguish between "actual" transitions where the user transitions in reality (at future time $t + K$), and "predicted" transitions where the user is "predicted" at time, $t + K$, to transition based on actual trajectory observations till $t + K - 1$. Consistent with our previous findings, we observe that $Model_{combidev}$ outperforms the baseline in both cases; we see approximately 35% improvement for $K = 1$ (i.e., in the next 15 minutes) and ≈20% with a 3 hour look-ahead time.

5.4.2 Non-overlapping Train/Test Time Series. As we deal with time series data in this work, a key concern during evaluation is the possibility of ground-truth leakage as a result of consecutive observation points from the time series becoming part of both train and test sets. This could potentially lead to an over-estimation of the performance observed.

In order to investigate this further, instead of randomly splitting the dataset to into 80-10-10% train-validation-test sets, we split the first half of the data (by date) into train and the remaining into equal parts of validation and test sets and re-ran the analysis. In Figure 13, we plot the resulting performance; we note that the performance remains relatively stable with only a ≈ 10% drop in performance for $K \in (1, 4)$. We also point out that this analysis was run on completely non-overlapping sets, although in practice, the performance should improve with online learning (i.e., a growing train set with each incoming test case and its corresponding prediction with some notion of confidence whose discussion we defer to future work).

5.4.3 Dynamic Predictions. Thus far, we have discussed the performance over the dataset covering the entire observation period of the data with mobility training of 4 weeks between 01-02-2017 through 28-02-2017, mobility predictions over the 2 week period of 01-03-2017 through 14-03-2017 which was then split into train/validation/test for non-conformance predictions. The tie strength values used thus far are the cumulative strengths calculated as at the end of this period.

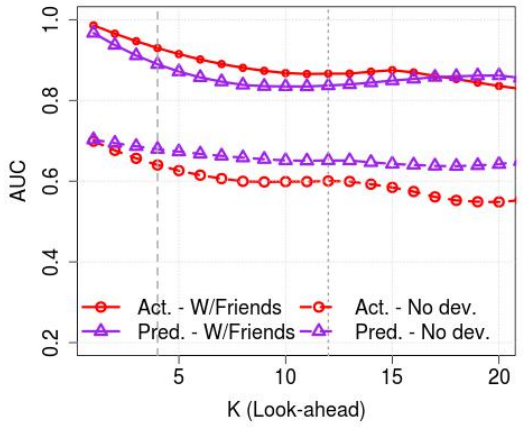


Fig. 12. Performance for (a) actual and (b) predicted instances of location transition.

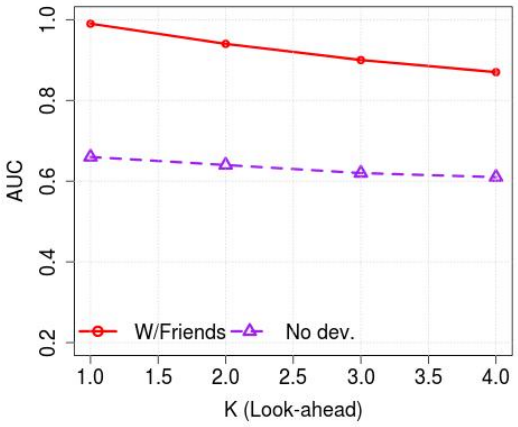


Fig. 13. Impact on Performance using Non-Overlapping Train/Test Time Series.

In order to understand how the performance would vary in practice where training data is acquired as the term progresses starting with zero data at the beginning of week 1 – i.e., the cold start problem, we study the online performance of week, w, using mobility training data from weeks, $[1, w-1]$, and tie strengths calculated as at the end week $w-1$. In Figure 14, we plot the performance of $Model_{combidev}$ and $Model_{nodev}$ for weeks 2 to 6. In both cases, we observe that the performance is relatively unstable for the first two weeks (with the least amount of training data) but that it stabilizes after week 4 with only marginal differences in performance beyond that period.

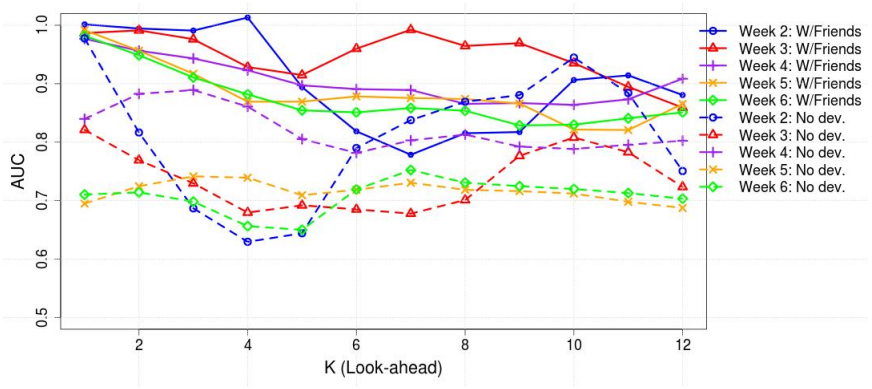


Fig. 14. Performance as the Term Progresses.

6 CASE STUDY: LOCATION-AWARE MOBILE CROWDSOURCING

There has been a significant body of research on the use of personal mobile devices to support various forms of participatory mobile sensing or crowdtasking in urban environments. A notable example of this paradigm

Table 4. Summary of LFNC Prediction Results.

	$K = 1$	$K = 2$	$K = 3$	$K = 4$
N_{sample}	114,470	113,544	112,316	111,036
AUC	1	0.96	0.93	0.9
Precision (@0.5)	99.11%	91.23%	86.42%	83.21%
Recall (@0.5)	99.11%	91.04%	86.16%	82.99%

is the use of campus users to report on the status of various campus resources/facilities (such as restrooms, cafeterias and office equipment) [8, 22, 23]. More specifically, a *Smart Campus* crowdtasking platform operational on-campus since 2014, uses the predicted movement pattern of participating users to recommend tasks that are likely to minimize a worker's detour overhead; empirical results show that this paradigm of crowdtasking based on trajectory predictions increases worker productivity by 60% [8]. In the crowd-tasking platform available on-campus, (i) the user's trajectory is derived based on identifying *staypoints* (the most likely location where the user spends the largest fraction of time within each 30 min window), and (ii) the task recommendations are made over distinct 3-hour windows. Empirical data also shows that 40% of tasks that are accepted by the platform workers are not eventually completed, with "unexpected" changes in the worker's movement pattern being cited as the most common cause for such non-completion. In such a scenario, the ability to better predict that the user is unlikely to be at specific predicted stay-points can be very valuable in improving the recommendation process.

In this section, we shall show that the use of our LFNC prediction can lead to significant gains in overall task completion rates and investigate its impact on worker productivity.

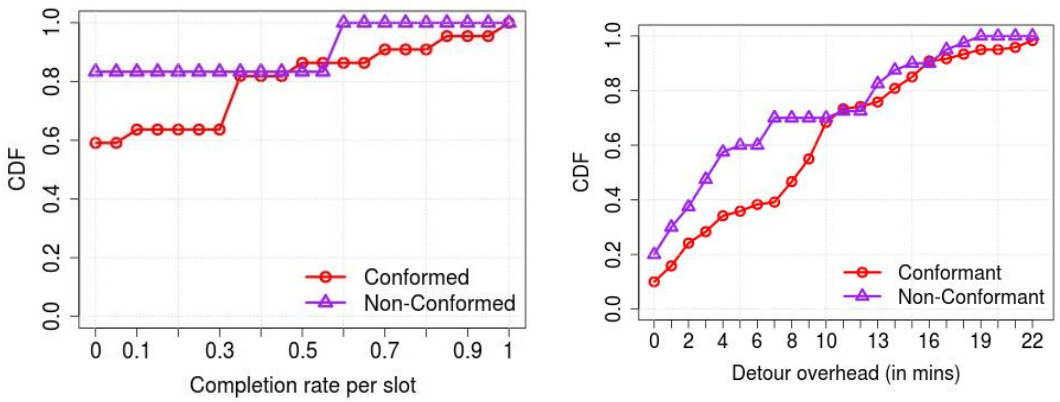
Crowd-tasking Data: A crowd-tasking pilot was carried out on-campus using the *Smart Campus* platform during a 2-week period of 14th March, 2017 through 31st March, 2017 (overlapping with *Dataset B*. A total of 325 student users were assigned tasks based on their predicted trajectories out of which 242 of them completed at least 1 task. Out of these 242, 106 (44%) of them were *recommended* tasks based on their historical movement behavior and predictions for the assigned task window – such an assignment is expected to minimize the student's detour overhead. In total, out of 60,000+ tasks assigned, and 3822 were completed with an overall task assignment-to-completion conversion rate of $\approx 6.2\%$.

6.1 LFNC Predictions

We utilize trajectory data of the *Smart Campus* users and their respective ego networks from 01-02-2017 to 13-03-2017 (i.e., X_{Train}) for mobility prediction training, and predict next place locations for $K-$ look-ahead distances over the pilot period. Out of this, as before in Section 5, we split the set into train/validation/test sets and make LFNC predictions over the validation and test tests. The predictions are then carried over as input to the task assignment module – we emphasize that, as this is a post-hoc analysis on an existing pilot, we are unable to assign tasks based on the LFNC predictions, but are only able to analyze differences in the two groups that were predicted to have been *conformant* and *nonconformant*. As the task assignments are made over 3-hour windows (3 times over the day at 9 AM, 12 Noon and 3 PM) where the students are allowed to perform the tasks any time during the window, we consider $K = 4$, i.e., predictions an hour ahead of the task time, and $1 \geq k \geq 5$, i.e., considering the student and his/her top-k ties' deviation where k depends on the number of ties who are present on campus concurrently.

6.2 Key Take-Aways

We first compare the detour incurred by students who were predicted to be *conformant* to their routine behaviour during the respective task window vs. those who weren't. In Figure 15b, we plot the detour, in minutes, the students from both categories incurred – the $y-$axis shows the CDF and $x-$axis represents the detour overhead.



(a) Completion Rate

(b) Detour/Overhead

Fig. 15. The difference in completion rates and detour incurred between the Tasker users whose behavior at task assignment time were predicted to be either "conformative" or "non-conformative", for $K = 4$ (i.e., with a look-ahead time of 1 hour) and $k = 2$.

For a student actual trajectory was from $A \rightarrow B$, and the student completes a task assigned to a location C, then the detour overhead is computed as $distance(A \rightarrow C) + distance(C \rightarrow B) - distance(A \rightarrow B)$. Here, the distance function returns the temporal distance, or the time taken to reach one location from the other. Interestingly, we see that the *nonconformant* group incurred statistically significant less detour in comparison to the *conformant* group ($D = 0.30833$, p-value = 0.006) – for instance, nearly 70% of tasks completed incurred ≤ 7 minutes of detour for the former whereas only 30% of the completed tasks incurred detours ≥ 7 minutes for the latter. We computed the entropy over the distribution of task locations that the two groups of workers chose and found that the *conformant* group showed a 12.5% increase compared to the *nonconformant* ($Entropy_{conformant} = 3.39$, $Entropy_{nonconformant} = 3.00$). In effect, the *non-comformant* group were unable to assist with tasks distributed throughout the campus, confining their task acceptance and execution to locations opportunistically close to their *actual* trajectories.

In Figure 15a, we plot the cumulative distribution function (CDF) on the $y-$axis and the average completion rate (over each unique user, time slot pair) on the $x-$axis (a student can be assigned multiple tasks during the same slot). We observe a statistically significant improvement in completion rates in the *conformant* group (represented by the red solid line) over *nonconformant* (blue dashed line) – a Kolmogorov Smirnov (KS) statistical test reveals a $p-$value of 0.0171 (i.e., < 0.05) and a D statistic of 0.476. In particular, we note that there is at least 20% improvement in the percentage of students who had an average completion rate of zero – whilst 80% of the nonconformant students had a zero completion rate, only 60% of the conformant students incurred the same. Also, we note a marginal increase in the *overall task completion rate* of 9% in the *conformant* group which is 3% more than the overall population. These results demonstrate the importance of such non-conformance prediction: a crowdsourcing platform aware of such users could choose to preferentially recommend tasks to other users, thereby increasing the overall task completion rate and the associated spatial diversity.

7 EXTENDING TO THE OUTDOOR SETTING

In the previous sections, we have described and evaluated our central hypotheses for predicting future non-conformance, in a predominantly-indoor urban campus. Here, we extend our analyses to an outdoor, city-scale

setting. Primarily, we hope to understand whether the inherent differences between indoor and outdoor mobility affect our capability to predict non-conformance.

Outdoor transit data: To study outdoor mobility, we exploit a public transit dataset from Singapore where each trip a commuter makes using the cashless payment card, on buses or trains, is captured along with the origin and destination station IDs and the corresponding timestamps. The dataset pertains to a period of 3 months from November, 2011 through January, 2012. In total, the dataset spans 300+ trips from over 5 million commuters, across 5000+ bus stops and train stations. For our analysis, we extract a set of 100-most frequent travelers (by total trip count on weekends over the period–see Figure 16), as well as their respective *co-travelers* (defined shortly, below). Whilst our indoor location data is updated periodically (every 2-3 minutes), the transit dataset is event-driven, containing location information only when a trip takes place. For the purpose of our analyses, we extrapolate the point-to-point trip data to construct trajectories (i.e., $x_{u,d}$, as defined in Section 2.2)– for instance, if a user enters $station_A$ at t_1, exits $station_B$ at t_2, re-enters $station_B$ at t_3, exits through $station_C$ at t_4, then the users taken to have *stayed* at $station_A$ and $station_B$ during t_1 to t_2 and t_2 to t_4, respectively.

Strength of Ties: As the social network information among commuters is also unavailable in this dataset, we adopt an approach similar to [20]. Trips that originate and terminate at the same stations within 20 seconds of each other at both entry and exit, during weekends, are considered to be *co-trajectories*, and the respective commuters considered to be *co-travelers*. For the 100-most frequent travelers, we extract such ego networks where the pair shares at least 2 co-trajectories. The strength of tie is then computed simply as the number of co-trajectories shared between the pair. We extract trajectories of 1024 travelers out of which 992 of them have taken a trip on at least 21 days. However, unlike in the case of the indoor dataset, we observe that a majority of travelers only have a single strong tie following this definition, and hence we limit our analysis to the *top-1* tie alone.

LFNC Prediction: As outdoor mobility is less frequent (longer stay duration) and outdoor location prediction is often at coarser granularity, we consider the locations of commuters at hourly intervals and at subzone level granularity[1]. More specifically, we map the geo-coordinates of the stations (the start and end points of a trajectory) to the corresponding subzone. To compute the *deviation* (i.e., the distance between the actual and expected trajectory), we sum up the Haversine distance between the corresponding locations in the two trajectories. Figure 17 plots the performance (AUC) for $Model_{nodev}$ and $Model_{combidev}$ with $k = 1$. Similar to our findings from the previous sections, we find that the deviations (from their normal routes), experienced by a commuter and the single strong tie, prove to be a reasonable early indicator of impending non-conformance – for instance, an $AUC \geq 0.85$ is observed for $K = 2$ hours. This represents a significant (30%) improvement in AUC over the deviation-unaware baseline, thereby demonstrating the power of our method.

8 DISCUSSION

K^{th}-likely Next Place Prediction: We have presently focused only on identifying non-conformance–i.e., in making a binary declaration of whether a user will visit the highest predicted location or not. By itself, this does not directly answer the question: where is the user most likely to be instead? As a plausible alternative, we can consider an expanded range of *top-K* ($K \geq 2$) predicted locations, and identify an anomaly only if the user does not visit any of these K locations. It is likely that such anomalies represent *dramatic disruptions* to the user's regular mobility pattern–e.g., a special annual concert on campus. It is unclear whether our hypotheses of 'temporal correlation of non-conformance' and 'homophily of anomalies' are valid for such rarer anomalies.

Other Applications and Alternate Anomaly Metrics: We believe that the ability to predict upcoming episodes of anomalous movement behavior can benefit many ubiquitous computing use cases, beyond the mobile crowd-sourcing application studied here. One such example is in dynamic calendaring applications, which can suggest

[1] https://data.gov.sg/dataset/master-plan-2014-subzone-boundary-web

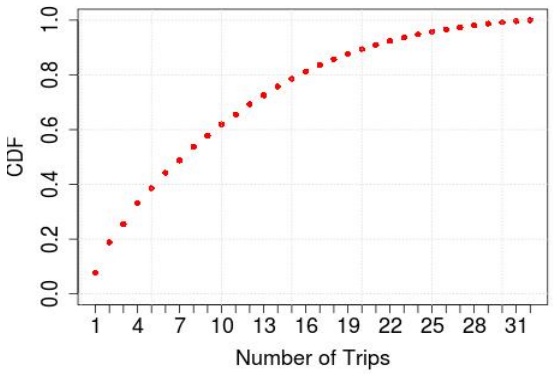


Fig. 16. Distribution of number of *weekend* trips taken by users during the 2-month observation period.

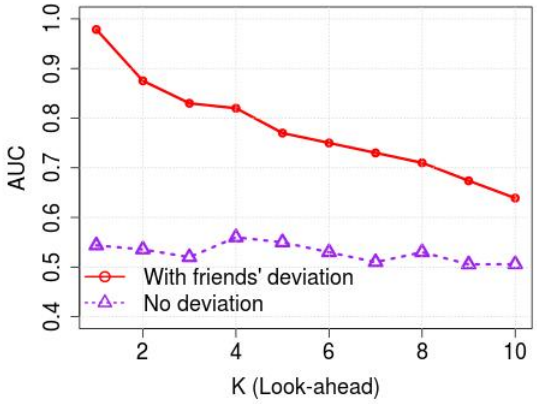


Fig. 17. LFNC Performance with increasing K (in hours).

Table 5. Summary of LFNC Prediction Results using Different Classification Algorithms.

Classifier	$K = 1$			$K = 2$			$K = 3$			$K = 4$		
	AUC	Precision	Recall	AUC	Precision	Recall	AUC	Precision	Recall	AUC	Precision	Recall
Decision Tree	0.809	0.863	0.8	0.812	0.861	0.815	0.791	0.828	0.795	0.776	0.801	0.779
Naive Bayes	0.832	0.794	0.78	0.828	0.791	0.779	0.806	0.768	0.761	0.791	0.758	0.751
Logistic Regression	0.841	0.86	0.812	0.843	0.857	0.815	0.82	0.821	0.794	0.805	0.795	0.777
Random Forest	0.843	0.853	0.813	0.845	0.852	0.816	0.827	0.82	0.796	0.814	0.795	0.779

schedule adjustments based on the attendance likelihood of participants. Another such example is *smart building energy management*, where predictions on the likely non-occurrence of regular meetings can help reduce energy consumption via proactive HVAC control techniques [10]. Both these cases, however, require more careful prediction of *collective LFNC* to determine the odds that multiple individuals will concurrently deviate from their normal mobility patterns.

Choice of Machine Learning Algorithm: All of our results presented to date use the GBM classifier. To study whether our insights on LFNC prediction are robust to the choice of classification technique, we conducted experiments (using the campus indoor mobility dataset) with additional shallow classifiers. Table 5 summarizes the key results for 4 different values of K (prediction window ranging from 15 mins—1 hour). We see that Logistic Regression and Random Forest classifiers seem to perform slightly better than the alternatives. However, the GBM classifier performs the best, achieving AUC of 0.9 for $K = 1$ (see Figure 8).

9 RELATED WORK

We first provide brief surveys of the body of knowledge on predictability of mobility and next place prediction from the literature, and then describe briefly, recent work that describe the interdependence of social ties and mobility, as well pointers towards context-aware crowd-tasking systems, which we use as a case study in this work.

Predictability of Human Mobility: While the ability to predict where a user will be next has many potential applications, works such as those of Song et al. [33], Lu et al. [24] and Jensen et al. [21] have focused on quantifying the theoretical bounds of the predictability of human mobility. If mobility is not intrinsically *regular* or *predictable*,

then the performance of next place prediction algorithms, however complex, will be limited. A natural choice for measuring randomness is entropy – Song et al. [33] use hourly mobility records of over 50,000 users to quantify (1) S_{rand} which is the entropy computed considering only number of different locations a user associates with, (2) S_{unc} which computes the entropy of a distribution over the different locations a user has been to and their associated frequency of visits, and finally, (3) S_{real} which considers both the frequency of visits and the sequence in which the locations were visited. S_{real} is computed as the Lempel-Ziv compression with the length tends to infinity. Using Fano's inequality, the authors derive an upper limit Π_{max} where Π is the probability of guessing the next location correctly with any algorithm. Further still, they discuss a lower bound R which the probability of the most likely location of a user during an hour of the day over a week. The authors make the key observation that mobility, taking into account past observations of frequency and sequence of visits, has very high median predictability (i.e., 93%) with significantly less variability across users, compared to using temporally uncorrelated location history. The authors investigate further the influence the distance traveled by users (through the radius of gyration) and demographics such as age and gender on predictability. Interestingly, even for users with high travel distances, the predictability remains high. While Song et al. [33] investigated outdoor mobility, Jensen et al. [21] study both outdoor (i.e., GSM and GPS) and indoor (i.e., via WLAN associations) by instrumenting the smartphones carried by 14 participants in Denmark. Whilst the GSM and WLAN records were sampled at faster rates (e.g., minutely), GPS, due to its large energy drain was sampled only 2-3 times every hour. Following from Song et al.'s work [33], the authors quantify the maximum predictability of mobility both indoors and outdoors; they find that even though the peak maximum predictability indoors is comparable to that of outdoor mobility, the variability across users is high. A key observation however is that the authors look at indoor mobility at the WLAN association level, and not at a *location* around it, whereas for many meaningful applications, localization to up to room level (e.g., 6- 8 meters) is sufficient. Hence, in our work we explore the modeling of mobility at the room level and floor level. The authors also explore different time scales and find that 3-4 minutes gives the best performance although the reason for this is not well-justified. More recently, Lu et al. [24] explore the use of Markov chains of increasing orders to investigate whether the theoretical maximum predictability is achievable practically. They rely on another large scale, outdoor mobility data set to show that Markov Chains of order 2 reach comparable accuracies to the theoretical maximum, and further note that the accuracies surpass the theoretical maximum for non-stationary trajectories (identified using the Gewek diagnostic). However, in this work, the authors consider a loose definition of a user's trajectory where they consider only the last recorded location of a user as the user's location for the day and the trajectory being composed of daily locations – we believe that the achievable predictability with such a definition whilst high, would have reduced practical benefits.

Next Place Prediction: Many works in wireless systems have investigated the practicality of predicting the next cell or location a user or mobile device is likely to be next. One of the earliest works was Reality Mining [13] with about 50 users where high predictability was reported using an order 2 Markov Chain to predict the next location over a limited semantic set (i.e., Home, Work or Other). Later, Kotz et al. [34] conduct a large-scale study of over 6000 students on the Dartmouth campus with observations from over 2 years. In this work, the authors compare the next cell prediction accuracy of two families of predictors: Markov Chains and Lempel-Ziv compressor based, and note down several key observations. Overall, they find that the added complexity of the LZ-family of predictors does not necessarily afford higher accuracy, and simple enhancements such as falling back to less complex models when past history does not contain the current context and accounting for recency can improve performance marginally. They also note that the accuracy is high only for users with long enough trajectories which might affect the practicality of such predictive algorithms. However, in this work, the authors only consider location changes as part of the user's trajectory and do not account for timing information – which again is a key attribute for practical applications.

The Next Place Prediction problem has been studied extensively due to its multitude of applications, but mostly in the context of outdoor movement derived from GPS traces from smartphones, taxicabs and social

media check-ins. Noulas et al. [28] study coarse-grained next place prediction using check-in data from over 1 million Foursquare (a popular Location-based Social Network platform) users where they consider transitions between categories of places, mobility flow between individual venues and share insights from spatio-temporal characteristics of check-in patterns. Gambs et al. [15] investigate the ability of Markov chains of order n to predict the next place of users, both indoors, using a phonetic dataset consisting of voice traces of 6 users, and outdoors, using the *GeoLife* dataset [38] consisting of GPS traces from Shanghai. Further, Baumann et al. [5] evaluate the problem extensively using 18 algorithms and their combination (using majority voting) on the Lausanne/Nokia MDC data set[2]. They report that although the accuracy is typically high, most errors are encountered during transitions from one place to the next. Spawning off from the Lausanne/Nokia MDC challenge, Gomes et al. [16], in addition to considering spatio-temporal history of traces, the authors augment contextual information accrued through sensors such as accelerometer, bluetooth and call/sms logs for better prediction. Further, Do et al. [12] discuss a variation of the problem, the probability of being at a specific location at a time in future using a dataset consisting of 133 smartphone users where they use kernel density estimation accounting additionally for day of the week and weekday/weekend effects. Our work uses similar methodologies such as those discussed in previous work in identifying possible next places, but is different in that the goal is in predicting, with sufficient look-ahead time, the possibility of the default next places predictions be incorrect – in other words, the problem reduces to providing a confidence measure of the predictions based on the current trajectory of a user. Separately, in Koehler et al. [**?**], the authors study the problem in two folds: (1) will the user stay at the current location for the next m time (i.e., temporally), and (2) if no, where will the user transition to next (i.e., spatially) using a number of machine learning techniques.

Mobility and Social Interactions: The expansive growth of Location-based Social Networks (LBSNs) such as Foursquare, and other popular mediums such as Twitter that allow for geo-tagging of posts, has led to many large-scale studies on urban mobility. The additional information declared by the users of the platform through features such as *follows* and explicit bidirectional friendships, makes it possible to infer the social relationships of the users, both offline and online. A number of works have focused on understanding the impact of such relationships on a user's mobility [9, 37]. Using physical trajectory data along with shared social relationship information, De Domenico et al.[11] report that incorporating knowledge of one's friend's mobility can help improve prediction of a user's mobility behavior. Recent works such as [7, 25] explored the use of body-worn social badges to infer and quantify face-to-face interactions of users in working environments. In these works, either friendship information is explicitly shared, or the participants are required to wear/carry additional sensors. In this work, we focus on inferring social ties *passively* using systems such as those described in past work [19, 20], and then utilize such social data-infused mobility information for predicting uncertainty in mobility behaviour.

Mobile Crowdsourcing: Location-aware mobile crowdsourcing has recently been employed to support the execution variety of reporting-centric tasks across both indoor and outdoor environments. The Ta$ker mobile crowdsourcing platform [23] uses predicted location trajectories, of university students to proactively recommend tasks that minimize a worker's travel detour. The user trajectories are computed as a series of staypoints–i.e., places where the user resides for large time periods. Using this platform, Kandappu et al. [22] showed that unexpected changes in trajectory caused around 6% of workers to "cheat"–i.e., report on tasks even though they did not visit the task location. Thus, LFNC prediction is useful not just in reducing unnecessary worker detour, but also in improving the reliability of responses.

[2]https://www.idiap.ch/dataset/mdc

10 CONCLUSIONS AND FUTURE WORK

Through our work, we have shown that the so-called 'random changes' in a user's routine movement behavior can, in fact, be predicted sufficiently in advance. For our empirical studies, we utilized a 4+ month WiFi-based indoor location dataset as our primary data source, and supplement it with an outdoor public transit data set. Empirically, our LFNC predictor (which uses both past anomalies and anomaly among strong-ties as key features) achieves a 35% improvement in prediction accuracy (a near-perfect AUC of 0.999) with a look-ahead time of 15 mins, and continues to outperform state-of-the-art Markovian mobility predictors even for longer look-ahead times (AUC =0.85 for K=2 hours). Moreover, on the campus data, the LFNC predictor's performance stabilizes after a modest observational period of 2-3 weeks.

We believe that this ability to accurately predict a user's divergence from her routine movement behavior has many practical applications. For example, crowdsourcing applications can avoid assigning or recommending tasks to users at risk of such mobility deviations, while smart building applications can perform proactive resource management to adapt to such likely anomalies (at an aggregated user level). In future work, we plan to embed our LFNC predictor's inputs in live deployments of such representative ubiquitous computing applications, and quantify the resulting performance gains. We also plan to investigate how user interfaces for commonplace workplace productivity applications (e.g., calendaring) can incorporate such "anomaly alerts".

ACKNOWLEDGMENTS

This material is supported partially by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centers in Singapore Funding Initiative. K. Jayarajah's work was supported by an A*STAR Graduate Scholarship. We thank Randy Tandriansyah Daratan for providing the data for the crowd-tasking analyses.

REFERENCES

[1] [n. d.]. Generalized Boosted Regression Model. https://cran.r-project.org/web/packages/gbm/index.html. ([n. d.]). [Online; Last accessed 15-Aug-2018].

[2] [n. d.]. ROCR: Visualizing the Performance of Scoring Classifiers. https://cran.r-project.org/web/packages/ROCR/index.html. ([n. d.]). [Online; Last accessed 15-Aug-2018].

[3] Paramvir Bahl and Venkata N Padmanabhan. 2000. RADAR: An in-building RF-based user location and tracking system. In *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, Vol. 2. Ieee, 775–784.

[4] Paul Baumann. 2014. Adaptive Sensor Cooperation for Predicting Human Mobility. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct)*.

[5] Paul Baumann, Wilhelm Kleiminger, and Silvia Santini. 2013. The influence of temporal and spatial features on the performance of next-place prediction algorithms. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 449–458.

[6] Richard A Becker, Ramon Caceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. 2011. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing* 10, 4 (2011), 18–26.

[7] Chloë Brown, Christos Efstratiou, Ilias Leontiadis, Daniele Quercia, Cecilia Mascolo, James Scott, and Peter Key. 2014. The architecture of innovation: Tracking face-to-face interactions with ubicomp technologies. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 811–822.

[8] Cen Chen, Shih-Fen Cheng, Aldy Gunawan, Archan Misra, Koustuv Dasgupta, and Deepthi Chander. 2014. Traccs: a framework for trajectory-aware coordinated urban crowd-sourcing. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

[9] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and Mobility: User Movement in Location-based Social Networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*.

[10] Paul Davidsson and Magnus Boman. 2005. Distributed Monitoring and Control of Office Buildings by Embedded Agents. *Inf. Sci.* 171, 4 (2005), 293–307.

[11] Manlio De Domenico, Antonio Lima, and Mirco Musolesi. 2013. Interdependence and Predictability of Human Mobility and Social Interactions. *Pervasive Mob. Comput.* 9, 6 (Dec. 2013).

[12] Trinh Minh Tri Do and Daniel Gatica-Perez. 2014. Where and what: Using smartphones to predict next locations and applications in daily life. *Pervasive and Mobile Computing* 12 (2014), 79–91.

[13] Nathan Eagle and Alex Pentland. 2006. Reality mining: sensing complex social systems. *Personal and ubiquitous computing* 10, 4 (2006), 255–268.

[14] Vincent Etter, Mohamed Kafsi, Ehsan Kazemi, Matthias Grossglauser, and Patrick Thiran. 2013. Where to go from here? Mobility prediction from instantaneous information. *Pervasive and Mobile Computing* 9, 6 (2013), 784–797.

[15] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. 2012. Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*. ACM, 3.

[16] João Bártolo Gomes, Clifton Phua, and Shonali Krishnaswamy. 2013. Where will you go? mobile data mining for next place prediction. In *International Conference on Data Warehousing and Knowledge Discovery*. Springer, 146–158.

[17] Richard W Hamming. 1950. Error detecting and error correcting codes. *Bell System technical journal* 29, 2 (1950), 147–160.

[18] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. 2011. Identifying important places in peopleâĂŹs lives from cellular network data. In *International Conference on Pervasive Computing*. Springer, 133–151.

[19] Kasthuri Jayarajah, Youngki Lee, Archan Misra, and Rajesh Krishna Balan. 2015. Need Accurate User Behaviour?: Pay Attention to Groups!. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*.

[20] K. Jayarajah, A. Misra, X. W. Ruan, and E. P. Lim. 2015. Event detection: Exploiting socio-physical interactions in physical spaces. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

[21] Bjørn Sand Jensen, Jan Larsen, Lars Kai Hansen, Jakob Eg Larsen, and Kristian Jensen. 2010. Predictability of mobile phone associations. In *Inter. Workshop on Mining Ubiquitous and Social Environments*.

[22] Thivya Kandappu, Nikita Jaiman, Randy Tandriansyah, Archan Misra, Shih-Fen Cheng, Cen Chen, Hoong Chuin Lau, Deepthi Chander, and Koustuv Dasgupta. 2016. TASKer: Behavioral Insights via Campus-based Experimental Mobile Crowd-sourcing. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*.

[23] Thivya Kandappu, Archan Misra, Shih-Fen Cheng, Nikita Jaiman, Randy Tandriansyah, Cen Chen, Hoong Chuin Lau, Deepthi Chander, and Koustuv Dasgupta. 2016. Campus-Scale Mobile Crowd-Tasking: Deployment & Behavioral Insights. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*.

[24] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson. 2013. Approaching the Limit of Predictability in Human Mobility. *Scientific Reports* 3 (Oct. 2013).

[25] Alessandro Montanari, Zhao Tian, Elena Francu, Benjamin Lucas, Brian Jones, Xia Zhou, and Cecilia Mascolo. 2018. Measuring Interaction Proxemics with Wearable Light Tags. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1 (March 2018).

[26] Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion* (2007), 69–84.

[27] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of Jaccard coefficient for keywords similarity. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Vol. 1.

[28] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. 2012. Mining user mobility features for next place prediction in location-based services. In *Data mining (ICDM), 2012 IEEE 12th international conference on*. IEEE, 1038–1043.

[29] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. 2007. Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences* 104, 18 (2007), 7332–7336.

[30] Eric Sven Ristad and Peter N Yianilos. 1998. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 5 (1998), 522–532.

[31] James Scott, A.J. Bernheim Brush, John Krumm, Brian Meyers, Michael Hazas, Stephen Hodges, and Nicolas Villar. 2011. PreHeat: Controlling Home Heating Using Occupancy Prediction. In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp '11)*. ACM, 281–290.

[32] Ivana Semanjski and Sidharta Gautama. 2015. Smart city mobility applicationâĂŤgradient boosting trees for mobility prediction and analysis based on crowdsourced data. *Sensors* 15, 7 (2015), 15974–15987.

[33] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of Predictability in Human Mobility. *Science* 327, 5968 (2010), 1018–1021. arXiv:http://science.sciencemag.org/content/327/5968/1018.full.pdf

[34] Libo Song, David Kotz, Ravi Jain, and Xiaoning He. 2003. Evaluating Location Predictors with Extensive Wi-Fi Mobility Data. *SIGMOBILE Mob. Comput. Commun. Rev.* 7, 4 (Oct. 2003), 64–65.

[35] Joe Tullio. 2003. Intelligent groupware to support communication and persona management. In *Proc. UIST 2003*.

[36] Joe Tullio, Jeremy Goecks, Elizabeth D Mynatt, and David H Nguyen. 2002. Augmenting shared personal calendars. In *Proceedings of the 15th annual ACM symposium on User interface software and technology*. ACM, 11–20.

[37] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. 2011. Human Mobility, Social Ties, and Link Prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*.

[38] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. 2009. Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th international conference on World wide web*. ACM, 791–800.

[39] Mengyu Zhou, Minghua Ma, Yangkun Zhang, Kaixin SuiA, Dan Pei, and Thomas Moscibroda. 2016. EDUM: classroom education measurements via large-scale WiFi networks. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 316–327.