Mobility Prediction Using Non-Parametric Bayesian Model

Jaeseong Jeong†, Mathieu Leconte‡ and Alexandre Proutiere†

Abstract—Mobility predictors proposed so far exploit the patterns from the past mobility trajectory collected by an individual user. Thus, if the training data collected by each user (i.e., the past trajectory) is not sufficient due to short training period or energy constraint, those predictors suffer from the low prediction accuracy. In this paper, we develop cluster-aided mobility predictors which use the data collected by all users to predict the next location of a given user. More specifically, based on a non-parametric bayesian statistics in which a clustering technique is involved, the proposed predictor exploits the data of other users whose mobility patterns are similar (i.e., associated to a same cluster). We analytically prove the consistency of the proposed predictor, and evaluate its performance using real mobility datasets.

I. INTRODUCTION

Predicting human mobility has received a great deal of attention recently, strongly motivated by a wide range of applications. Examples of such applications include: location-based services provided to users by anticipating their movements (e.g., mobile advertisement, recommendation systems, risk alarm); urban traffic engineering and forecasting; the design of more efficient radio resource allocation protocols in wireless networks (e.g., scheduling and handover management [1], data prefetching [2] and energy efficient location sensing [3]). However, the improvement of these applications achieved by predicting mobility critically depends on the accuracy of the delivered predictions.

The trajectories of users in networks contain several typical and repeated patterns, which may be exploited to predict future mobility. These patterns correspond to the regular behaviors of users, e.g., commuting from home to work or visiting favourite restaurants. Identifying such patterns in the mobility history of a user is key to making good predictions of her future mobility. Unfortunately, there are many challenges associated with gathering histories of past mobility. For instance, detecting the current location of a user with sensors (e.g., GPS, Wi-Fi and cell tower) consumes a non-negligible energy. Users may also hesitate to log their trajectories due to privacy issues. If the gathered history of the mobility of a user is not sufficient due to above reasons, her typical mobility patterns may not be identified or may contain a lot of noise, which will in turn result in poor mobility predictions.

Many location predictors have been devised over the last decade [4–8]. The predictors proposed so far estimate the next location of a specific user by inspecting the past individual trajectories of this user. One of the most popular mobility predictors consists in modelling the user trajectory as an order-κ Markov chain. Predictors based on the order-κ Markov model are asymptotically optimal [7], [8] for a large class of mobility models, provided that the order κ slowly grows with the length of the observed mobility history. This optimality only holds asymptotically when the length of the observed user past trajectory tends to infinity. Unfortunately, when the past trajectory of the user is rather short, these predictors perform poorly.

In this paper, we aim at devising mobility predictors that perform well even if the past trajectories gathered for the various users are short. Our main idea is to develop cluster-aided predictors that exploit the data (i.e., past trajectories) collected from all users to predict the next location of a given user. These predictors rely on clustering techniques and extract from the training data similarities among the mobility patterns of the various users to improve the prediction accuracy. More precisely, we make the following contributions:

• We present CAB (Cluster-Aided Bayesian), a cluster-aided predictor whose design is based on recent non-parametric bayesian statistical tools [9], [10]. CAB extracts from the data clusters of users with similar mobility processes, and exploit this clustered structure to provide accurate mobility predictions. The use of non-parametric statistical tools allows us to adapt the number of extracted clusters to the training data (this number can actually grow with the data). This confers to our algorithm a strong robustness, i.e., CAB exploits similarities in users’ mobility only if such similarities are really present in the training data.

• We derive theoretical performance guarantees for the CAB predictor. In particular, we show that CAB is asymptotically optimal (among the set of all predictors) when the number of users grows large, and for a large class of mobility models.

• Using simulations, and artificially generated user trajectories, we illustrate the superiority and the aforementioned robustness of the CAB predictor.

• Finally, we compare the performance of our predictor to that of other existing predictors using two large-scale mobility datasets (corresponding to a WiFi and a cellular network, respectively). CAB significantly outperforms existing predictors, and in particular those that only exploit individual past trajectories to estimate users’ next location.
II. RELATED WORK

Most of existing mobility prediction methods consist in estimating the next location of a specific user by inspecting the past individual trajectories of this user. The order-$k$ Markov predictor [4] is known to be optimal [11] when the trajectory is generated by a Markov process. In order to improve the performance of this predictor for small histories, a fallback mechanism can also be added [4] to reduce the order of the Markov model when the current sequence of $k$ previous locations has not been encountered before. In order to adapt the order of the Markov model used for prediction, [8] proposes Sampled Pattern Matching (SPM), which sets the order of the Markov model to a fraction of the longest suffix match in the history. This predictor is asymptotically optimal with provable bounds on the rate of convergence, while the trajectory is generated by a stationary mixing source. Another predictor with fallback gives comparable performance to SPM [8]. NextPlace [5] performs a non-linear time-series analysis on the history to predict the next visiting time of a user to any of her frequently visited locations.

Empirical evaluations [4], [6] show that complex individual mobility models do not perform well: the order-2 Markov predictor with fallback gives comparable performance to SPM [8]. NextPlace [5] and higher order Markov predictors. In addition, [3], [6] report that the order-1 Markov predictor can actually provide better predictions than higher order Markov predictors, as the latter suffer more from the lack of training data.

As far as we know, this paper provides the first cluster-aided mobility prediction method. However, there have been a few papers aiming at clustering trajectories or more generally stochastic processes. For example, [12] proposes algorithms to find clusters of trajectories based on likelihood maximization for an underlying hidden Markov model. For the same problem, [13] uses spectral clustering in a semi-parametric manner based on Bhattacharyya affinity metric between pairs of trajectories. However, those methods would not work well in our setting. This is due to the facts that (i) users belonging to a same cluster should have trajectories generated by identical parameters, and (ii) the number of clusters should be known beforehand, or estimated in a reliable way. The proposed non-parametric Bayesian approach address both these issues.

III. MODELS AND OBJECTIVES

In this section, we first describe the data on past user trajectories available at a given time to build predictors of the next user locations. We then provide a simple model for user mobility, used to define our non-parametric inference approach, as well as its objectives.

A. Collected Data

We consider the problem of predicting at a given time the mobility, i.e., the next position, of users based on observations about past users’ trajectories. These observations are collected and stored on a server, also responsible for providing mobility predictions. The set of users is denoted by $\mathcal{U}$, and users are all moving within a common finite set $\mathcal{L}$ of $L$ locations. The observed trajectory for user $u$ is denoted as $x^u = (x^u_1, \ldots, x^u_{n^u})$, where $x^u_t$ corresponds to the $t$-th location visited by user $u$, and where $n^u$ refers to the length of the trajectory. $x^u_n$ denotes the current location of user $u$.

By definition, we impose $x^u_t \neq x^u_{t+1}$, i.e., two consecutive locations on a trajectory must be different. Let $x^t = (x^u)_{u \in \mathcal{U}}$ denote the collection of user trajectories. It is worth noting that our notion of trajectory does not involve the speed at which users move, but only considers the sequence of different visited locations. We believe that this notion constitutes an appropriate model for most existing mobility traces, where users do not necessarily report their location periodically. Also observe that the lengths of the trajectories may vary across users, which again seems natural.

We introduce a few additional notations. We denote by $n^u_{i,j}$ the number of observed transitions of user $u$ from location $i$ to location $j$, i.e., $n^u_{i,j} = \sum_{t=1}^{n^u-1} 1(x^u_t = i, x^u_{t+1} = j)$. Let $\mathcal{H} = \cup_{n=0}^{\infty} \mathcal{U}^n$ denote the set of all possible trajectories of a single user, and let $\mathcal{H}^\mathcal{U}$ be the set of all possible set of trajectories of users in $\mathcal{U}$.

B. Mobility Models

The design of our predictors is based on a simple mobility model. We assume that user trajectories are order-1 Markov chains, with arbitrary initial state or location. More precisely, user-$u$’s trajectory is generated by the transition kernel $\theta^u = (\theta^u_{i,j})_{i,j \in \mathcal{L}} \in [0,1]^{L \times L}$, where $\theta^u_{i,j}$ denotes the probability that user $u$ moves from location $i$ to $j$ along her trajectory. Hence, given her initial position $x^u_1$, the probability of observing trajectory $x^u$ is $P_{\theta^u}(x^u) := \prod_{t=1}^{n^u-1} \theta^u_{x^u_t,x^u_{t+1}}$. Our mobility model can be readily extended to order-$k$ Markov chains. However, as observed in [6], order-1 Markov chain model already provides reasonably accurate predictions in practice, and higher-order models would require a fallback mechanism [4]. Throughout the paper, we use uppercase letters to represent random variables and the corresponding lowercase letters for their realizations, e.g., $X^u$ (resp. $x^u$) denotes the random (resp. realization of) trajectory of user $u$.

C. Bayesian Framework, Clusters, and Objectives

In this paper, we adopt a Bayesian framework, and we assume that the transition kernels of the various users are drawn independently from the same distribution $\mu \in \mathcal{P}(\Theta)$ referred to as the prior distribution over the set of all possible transition kernels $\Theta$. This assumption is justified by the De Finetti’s theorem (see [14], Theorem 11.10) if $\theta_{u \in \mathcal{U}}$ are exchangeable (which is typically the case if users are a priori indistinguishable). In the following, the expectation and probability under $\mu$ are denoted by $\mathbb{E}$ and $\mathbb{P}$, respectively. To summarize, the trajectories of users are generated using the following hierarchical model: for all $u \in \mathcal{U}$, $\theta^u \sim P_{\theta^u}$, and $n^u, X^u_1$ are arbitrarily fixed.

1To accurately predict the next position of user $u$ given that the sequence of her past $k$ positions is $i_1, \ldots, i_k$, her trajectory should contain numerous instances of this sequence, which typically does not occur if the observed trajectory is short – and this is precisely the case we are interested in.
To improve the accuracy of our predictions, we leverage similarities among user mobility patterns. It seems reasonable to think that the trajectories of some users are generated through similar transition kernels. In other words, the distribution $\pi$ might exhibit a clustered structure, putting mass around a few typical transition kernels. Our predictors will identify these clusters, and exploit this structure, i.e., to predict the next location of a user, we shall leverage the observed trajectories of all users who belong to user-$u$'s cluster.

For any user $u$, we aim at proposing an accurate estimator (or predictor) $\hat{x}^u \in L$ of her next location, given the observed trajectories $X^{u|t} = x^{u|t}$ of all users. The Bayesian accuracy of a predictor $\hat{x}^u$ for user $u$, denoted by $\pi_u(\hat{x}^u)$, is defined as $\pi_u(\hat{x}^u) := \mathbb{P}(X^{u|n+1} = \hat{x}^u \mid x^{u|t}) = \mathbb{E}[\theta^{u}_{\varphi_{\pi_{n+1}}}] \mid x^{u|t}]$ (where for conciseness, we write $\mathbb{P}(\cdot \mid x^{u|t}) = \mathbb{P}(\cdot \mid X^{u} = x^{u|t})$). Clearly, given $X^{u|t} = x^{u|t}$, the best possible predictor would be:

$$\hat{x}^u \in \arg \max_{j \in L} \mathbb{E}[\theta^{u}_{j}] \mid x^{u|t}].$$

Computing this optimal predictor, referred to as the Bayesian predictor with prior $\mu$, requires the knowledge of $\mu$. Indeed:

$$\mathbb{E}[\theta^{u}_{j}] \mid x^{u|t}] = \int_{\varphi} \theta^{u}_{j} \mu(\varphi) d\varphi = \frac{\int_{\varphi} \theta^{u}_{j} \mu(\varphi) d\varphi}{\int_{\varphi} \mu(\varphi) d\varphi}.
$$

To circumvent this issue, we will first approximate the unknown distribution $\mu$, and then construct $\hat{\theta}^u$.

IV. BAYESIAN NONPARAMETRIC INFERENCE

In view of the model described in the previous section, we can devise an accurate mobility predictor if we are able to provide a good approximation of the prior distribution $\mu$ on the transition kernels dictating the mobility of the various users. If $\mu$ concentrates its mass around a few typical kernels that would in turn define clusters of users (i.e., users with similar mobility patterns), we would like to devise an inference method identifying these clusters. On the other hand, our inference method should not discover clusters if there are none, nor specify in advance the number of clusters (as in the traditional mixture modelling approach). Towards these objectives, we apply a Bayesian non-parametric approach that estimates how many clusters are needed to model the observed data and also allows the number of clusters to grow with the size of the data. In Bayesian non-parametric approaches, the complexity of the model (here the number of clusters) is part of the posterior distribution, and is allowed to grow with the data, which confers flexibility and robustness to these approaches.

In the remaining section of this paper, we first present an overview of the Dirichlet Process mixture model, a particular Bayesian non-parametric model, and then apply this parametric model to the design of CAB (Cluster-Aided Bayesian), a robust and flexible prediction algorithm that efficiently exploit similarities in users’ mobility, if any.

A. Dirichlet Process Mixture Model

When applying Bayesian non-parametric inference techniques to our prediction problem, we add one level of randomness. More precisely, we approximate the prior distribution $\mu$ on the transition kernels $\theta^u$ by a random variable, also referred to as $\mu$, with distribution $g \in \mathcal{P}(\mathcal{P}(\Theta))$. This additional level of randomness allows us to introduce some flexibility in the number of clusters present in $\mu$. We shall compute the posterior distribution $g$ given the observations $x^{u|t}$, and hope that this posterior distribution, denoted as $g \mid x^{u|t}$, will concentrate its mass around the true prior distribution $\mu$.

To evaluate $g \mid x^{u|t}$, we use Gibbs sampling techniques (see Section IV-B1), and from these samples, we shall estimate the true prior $\mu$, and derive our predictor by replacing $\mu$ by its estimate in (1-2).

For the higher-level distribution $g$, we use the Dirichlet Process (DP) mixture model, a standard choice of prior over infinite dimensional spaces, such as $\mathcal{P}(\Theta)$. The DP mixture model has a possibly infinite number of mixture components or clusters, and is defined by a concentration parameter $\alpha > 0$, which impacts the number of clusters, and a base distribution $G_0 \in \mathcal{P}(\Theta)$, from which new clusters are drawn. The DP mixture model with parameters $\alpha$ and $G_0$ is denoted by $DP(\alpha, G_0)$ and defined as follows. If $\nu$ is a random measure drawn from $DP(\alpha, G_0)$ (i.e., $\nu \sim DP(\alpha, G_0)$), and $\{A_1, A_2, \ldots, A_K\}$ is a (measurable) partition of $\Theta$, then $(\nu(A_1), \ldots, \nu(A_K))$ follows a Dirichlet distribution with parameters $(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))$. It is well known [15] that a sample $\nu$ from $DP(\alpha, G_0)$ has the form

$$\nu = \sum_{c=1}^{\infty} \beta^c \delta_{\theta^c},$$

where $\delta_{\theta}$ is the Dirac measure at point $\theta \in \Theta$, the $\delta_{\theta^c}$’s are i.i.d. with distribution $G_0$ and represent the centres of the clusters, and the weights $\beta^c$’s are generated using a Beta distribution according to the following stick-breaking construction:

$$\beta^c \sim \text{Beta}(1, \alpha), \text{ i.i.d.,}$$

$$\beta^c = \frac{\beta^c}{\sum_{i=1}^{c-1} \beta^i}.$$

When $(\theta^u)_{u \in U}$ is generated under the above DP mixture model, we can compute the distribution of $\theta^u$ given $\theta^\setminus u (\theta^\setminus u)$ is a (measurable) partition of $\Theta$. Then in $U \setminus \{u\}$, and $c \in \mathcal{E}(\setminus u)$ are clustered and the set of corresponding clusters is denoted by $c \setminus u$. Users in cluster $c$ in $\mathcal{E}(\setminus u)$ share the same transition kernel $\theta^c$, and the number of users assigned to cluster $c$ is denoted by $n_{c \setminus u} = \sum_{u \in U \setminus \{u\}} I_{u \in c \setminus u}$. The distribution of $\theta^u$ given $\theta^\setminus u$ is then:

$$\theta^u \mid \theta^\setminus u \sim \begin{cases} G_0 & \text{w.p. } \frac{\alpha}{\alpha + |\setminus u| - 1}, \\ \delta_{\theta^c} & \text{w.p. } \frac{|\setminus u| - 1}{\alpha + |\setminus u| - 1}, \forall c \in \mathcal{E}(\setminus u). \end{cases}$$

makes the cluster structure of the DP mixture model explicit. Indeed, when considering a new user $u$, a new cluster containing user $u$ only is created with probability $\frac{|\setminus u| - 1}{\alpha + |\setminus u| - 1}$, and user $u$ is associated with an existing cluster $c$ with probability proportional to the number of users already assigned to this cluster. Refer to [16] for a more detailed description on DP mixture models.

[1]The Dirichlet distribution with parameters $(\alpha_1, \ldots, \alpha_K)$ has density (with respect to Lebesgue measure) proportional to $\prod_{k=1}^{K} x^{\alpha_k}_k$.\[2\text{The Dirichlet distribution with parameters } (\alpha_1, \ldots, \alpha_K) \text{ has density (with respect to Lebesgue measure) proportional to } \prod_{k=1}^{K} x^{\alpha_k}_k.\]
Our prediction method simply consists in approximating \( \mathbb{E}[\theta^u|x^L] \) by the expectation w.r.t. the posterior distribution \( g|x^L \). In other words, for user \( u \), the estimated next position will be:

\[
\hat{x}^u \in \arg\max \mathbb{E}_{g}[\theta^u|z^u_{n+J}, x^L],
\]

where \( \mathbb{E}_g[\cdot] \) denotes the expectation w.r.t. the probability measure induced by \( g \). To compute \( \mathbb{E}_g[\theta^u|x^L] \), we rely on Gibbs sampling techniques to generate samples with distribution \( g|x^L \). The way \( g|x^L \) concentrates its mass around the true prior \( \mu \) will depend on the choices of parameters \( \alpha \) and \( G_0 \), and to improve the accuracy of our predictor, these parameters will be constantly updated when successive samples are produced.

### B. Cluster-Aided Bayesian (CAB) Predictor

Next we present CAB, our mobility prediction algorithm. The objective of this algorithm is to estimate \( \hat{\theta}^u \) from which we derive the predictions according to (5). CAB consists in generating independent samples of the assignment of user to clusters induced by the posterior distribution \( g|x^L \), and in deducing an estimate of \( \mathbb{E}_g[\theta^u|x^L] \). As mentioned above, the accuracy of this estimate strongly depends on the choice of parameters \( \alpha \) and \( G_0 \) in the DP mixture model, and these parameters will be updated as new samples will be generated.

More precisely, the CAB algorithm consists in two steps. (i) In the first step, we use Gibbs sampler to generate \( B \) samples of the assignment of users to clusters under the probability measure induced by \( g|x^L \), and update the parameters \( \alpha \) and \( G_0 \) of the DP mixture model using these samples. Hence we update the prior distribution \( g \). We repeat this procedure \( K \)-1 times. In the \( k \)-th iteration, we construct \( B \) samples of users’ assignment. The \( b \)-th assignment sample is referred to as \( \epsilon^L_{b,k} = (\epsilon_{a,b,k})_{a \in \mathcal{U}} \) in CAB pseudo-code, where \( \epsilon_{a,b,k} \) is the cluster of user \( u \) in that sample. The subroutines providing the assignment samples, and updating the parameters of the prior distribution \( g \) are described in details in [IV-B1] and [IV-B2] respectively. At the end of the first step, we have constructed a prior distribution \( g \) parametrized by \( G_0^k \) and \( \alpha_K \) which is adapted to the data, i.e., a distribution that concentrates its mass on the true prior \( \mu \). (ii) In the second step, we use the updated prior \( g \) to generate one last time \( B \) samples of users’ assignment. Using these samples, we compute an estimate \( \hat{\theta}^u \) of \( \mathbb{E}_g[\theta^u|x^L] \) for each user \( u \), and finally derive the prediction \( \hat{x}^u \) of the next position of user \( u \). The way we compute \( \hat{\theta}^u \) is detailed in [IV-B3].

The CAB algorithm takes as inputs the data \( x^L \), the number \( K \) of updates of the prior distribution \( g \), the number of samples \( B \) generated by the Gibbs sampler in each iteration, and the number of times \( M \) the users’ assignment is updated when producing a single assignment sample using Gibbs sampler (under Gibbs sampler, the assignment is a Markov chain, which we simulate long enough so as it has the desired distribution). \( K, B, \) and \( M \) have to be chosen as large as possible. Of course, increasing these parameters also increases the complexity of the algorithm, and we may wish to select the parameters so as to achieve a appropriate trade-off between accuracy and complexity.

#### Algorithm 1 CAB Predictor

**Input:** \( x^L, K, B, M \)

1) **Step 1: Updates of \( G_0 \) and \( \alpha \)**

\[
G_0 \leftarrow \text{Uniform}(\Theta), \alpha_k \leftarrow 1
\]

for \( k = 1 \ldots K - 1 \) do

for \( b = 1 \ldots B \) do

\[ c^L_{b,k} \leftarrow \text{GibbsSampler}(x^L, G_0^k, \alpha_k, M) \]

end

\[ G_0^{k+1}, \alpha_{k+1} \leftarrow \text{UpdateDP}(x^L, G_0^k, \{c^L_{b,k}\}_{b=1 \ldots B}) \]

end

2) **Step 2: Last sampling and prediction**

for \( b = 1 \ldots B \) do

\[ c^L_{b,K} \leftarrow \text{GibbsSampler}(x^L, G_0^K, \alpha_K, M) \]

end

Compute \( \hat{\theta}^u \) by implementing (5) using \( \{c^L_{b,K}\}_{b=1 \ldots B} \) and \( G_0^K \)

\[ \hat{x}^u = \arg\max_j \hat{\theta}^u_{z^u_{n+J}} \]

**Output:** \( \hat{\theta}^u, \hat{x}^u \)

1) **Sampling from the DP mixture posterior:** We use Gibbs sampler [17] to generate independent samples of the assignment of users to clusters under the probability measure induced by the posterior \( g|x^L \), i.e., samples of assignment with distribution \( P_g[c^L|x^L] \), where \( P_g \) denotes the probability measure induced by \( g \). Gibbs sampling is a classical MCMC method to generate samples from a given distribution. It consists in constructing and simulating a Markov chain whose stationary state has the desired distribution. In our case, the state of the Markov chain is the assignment \( c^L \), and its stationary distribution is \( P_g[c^L|x^L] \). The Markov chain should be simulated long enough (here the number of steps is denoted \( M \)) so that at the end of the simulation, the state of the Markov chain has converged to the steady-state. The pseudocode of the proposed Gibbs sampler is provided in Algorithm 2 and easily follows from the description of the DP mixture model provided in [3].

To produce a sample of the assignment of users to clusters, we proceed as follows. Initially, we group all users in the same cluster \( c_1 \), the number of cluster \( N \) is set to 1, and the number of users (except for user \( u \)) \( n_{c_1,u} \) assigned to cluster \( c_1 \) is \( |\mathcal{U}| - 1 \). (see Algorithm 2). Then the assignment is revised \( M \) times. In each iteration, each user is considered and assigned to either an existing cluster, or to a newly created cluster (the latter is denoted by \( c_{N+1} \) if in the previous iteration there was \( N \) clusters). This assignment is made randomly according to the model described in [4]. Note that in the definition of \( \beta_c \), we have \( G_0(d\beta|x^c) = \frac{P_g(x^c|G_0(\alpha))}{\int_{\Theta} P_g(x|G_0(\alpha))d\alpha} \), where \( x^c \) corresponds to the data of users in cluster \( c \), i.e., \( x^c = (x^u)_{u \in c} \).

2) **Updates of \( G_0 \) and \( \alpha \):** As in any Bayesian inference method, our prediction method could suffer from a bad choice of parameters \( \alpha \) and \( G_0 \) defining the prior \( g \). For example, by choosing a small value for \( \alpha \), we tend to get a very small number of clusters, and possibly only one cluster. On the contrary, selecting a too large \( \alpha \) would result in a too large number of clusters, and in turn, would make our algorithm unable to capture similarities in the mobility patterns of the various users. To circumvent this issue, we update and fit the
Algorithm 2 GibbsSampler

Input: $x^U, G_0, \alpha, M$
\forall u \in U, c_u \leftarrow c_1, n_{c,-u} \leftarrow |U| - 1; N \leftarrow 1; c^U = \{c_1\}.
for \( i = 1 \ldots M \) do
\hspace{1em} for each \( u \in U \) do
\hspace{2em} \( c^u \leftarrow c^u \setminus \{u\} \)
\hspace{2em} \( \beta_{new} \leftarrow z \left[ c_{new} \leftarrow \left\{ u \right\} \right] P_0(x^u)G_0(d\theta) \)
\hspace{2em} \( \beta_{c} \leftarrow z \left[ c_{N+1} \leftarrow \left\{ u \right\} \right] P_0(x^u)G_0(d\theta|x^c), \forall c \in c^u \setminus \{u\} \)
\hspace{2em} In the above expressions, \( z \) is a normalizing constant, i.e., selected so that \( \beta_{new} + \sum_{c \in c^u \setminus \{u\}} \beta_{c} = 1 \).
\hspace{2em} With probability \( \beta_{new} \) do:
\hspace{3em} \( c^u \leftarrow c_{N+1}; c_{N+1} \leftarrow \{u\}; n_{c_{N+1}, -u} \leftarrow 0; n_{c_{N+1}, c} \leftarrow 1, \forall v \neq u; c^U \leftarrow c^U \cup \{c_{N+1}\}; N \leftarrow N + 1; \)
\hspace{2em} and with probability \( \beta_{c} \) do:
\hspace{3em} \( c^u \leftarrow c; c \leftarrow c \cup \{u\}; n_{c, -v} \leftarrow n_{c, -v} + 1, \forall v \neq u. \)
end
end
Output: Cluster assignment \( c^U \)

parameters to the data, as suggested in \cite{10}. In the CAB algorithm, the initial base distribution is uniform over all transition kernels (over \( \Theta \)) and \( \alpha \) is taken equal to 1. Then after each iteration, we exploit the samples of assignments of users to clusters to update these initial parameters, by refining our estimates of \( G_0 \) and \( \alpha \).

Algorithm 3 UpdateDP at the \( k \)-th iteration

Input: \( x^U, G^K_0, \{c^{U,b,k}\}_{b=1 \ldots B} \)
Compute \( G_0^{k+1}(\cdot) \) and \( \alpha_{k+1} \) as follows.
\[ G_0^{k+1}(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \sum_{c^{U,b,k}} \frac{n_{c^{U,b,k}}}{|U|} G^K_0(\cdot|x^c) \]
\[ \alpha_{k+1} = \arg \min_{\alpha_{k}} \left\{ \sum_{i=1}^{B} \left[ \frac{|U|}{\alpha + i - 1} - \frac{1}{B} \sum_{b=1}^{B} N_b \right] \right\} \]
where \( n_{c^{U,b,k}} \) is the size of cluster \( c \in c^{U,b,k} \), and \( N_b \) is the total number of (non-empty) clusters in \( c^{U,b,k} \).

Output: \( G_0^{k+1}, \alpha_{k+1} \)

Note that (6) simply corresponds to a kernel density estimator based on the \( B \) cluster samples obtained with prior distribution parametrized by \( G^K_0 \) and \( \alpha \), whereas (7) corresponds to a maximum likelihood estimate (see \cite{13}), which sets \( \alpha_{k+1} \) to the value which is most likely to have resulted in the average number of clusters obtained when sampling from the model with parameters \( G^K_0 \) and \( \alpha_k \).

3) Computation of \( \hat{\theta}^u \): As mentioned earlier, \( \hat{\theta}^u \) is an estimator of \( E_g(\hat{\theta}^u|x^U) \), where \( g \) is parametrized by \( G^K_0 \) and \( \alpha_K \), and is used for our prediction of user-\( u \)’s mobility. \( \hat{\theta}^u \) is just the empirical average of \( \hat{\theta}^c \) for clusters \( c \) to which user-\( u \) is associated in the \( B \) last samples generated in CAB, i.e.,
\[ \hat{\theta}^u = \frac{1}{B} \sum_{b=1}^{B} E_g(\hat{\theta}^{u,b,k} | x^{u,b,k}) \]
\[ = \frac{1}{B} \sum_{b=1}^{B} \left[ \frac{\theta \cdot P_0(x^{u,b,k})G^K_0(d\theta)}{\int \theta \cdot P_0(x^{u,b,k})G^K_0(d\theta)} \right] \]
Note that in view of the law of large numbers, when \( B \) grows large, \( \hat{\theta}^u \) converges to \( E_g(\hat{\theta}^u|x^U) \). The predictions for user \( u \) are made by first computing an estimated transition kernel \( \hat{\theta}^u \) according to (9). We derive an explicit expression of \( \hat{\theta}^u \) that does not depend on \( G^K_0 \), but only on data and the samples generated in the CAB algorithms. This expression in the following lemma will be useful to understand to what extent the prediction of user-\( u \)’s mobility under CAB leverages observed trajectories of other users.

Lemma 1 For any \( i, j \), \( \hat{\theta}^u_{i,j} \) is computed by a weighted sum of all users’ empirical transition kernels \( (n^u_{i,j}/n^u_v, v \in U) \), i.e.,
\[ \hat{\theta}^u_{i,j} = \sum_{c_{i}, \ldots, c_{K}} \xi_{c_{i}, \ldots, c_{K}} \frac{1}{|L| + \sum_{k=1}^{K} n^u_{c_{K}}} \prod_{k=1}^{K} \omega^{c_{K}}_{k} + \]
\[ \sum_{v \in U} \sum_{c_{i}, \ldots, c_{K}} \xi_{c_{i}, \ldots, c_{K}} n^v_{i,j} \prod_{k=1}^{K} \omega^{c_{K}}_{k} \prod_{k=1}^{K} \omega^{c_{K}}_{k} \]
where the sum \( \sum_{c_{i}, \ldots, c_{K}} \) is over all \( c_{i}, \ldots, c_{K} \) and \( \omega^{c_{K}}_{k} \) is a set of every cluster sampled at \( k \)-th iterations (i.e., \( \omega^{c_{K}}_{k} = \{c \mid \sum_{b=1}^{B} \sum_{u \in U} \mathbb{1}(c^{U,b,k} = c) > 0 \} \)). \( \omega^{c_{K}}_{k} \) and \( \xi_{c_{i}, \ldots, c_{K}} \) are
\[ \xi_{c_{i}, \ldots, c_{K}} = \prod_{i \in L} \Gamma(1 + \sum_{k=1}^{K} n^u_{c_{K}}) \]
\[ \omega^{c_{K}}_{k} = \frac{n^u_{c_{K}}}{B|L| \sum_{c_{i}, \ldots, c_{K-1}, c} \prod_{k=1}^{K-1} \omega^{c_{K}}_{k}} \]
where \( n^u_{c_{i} \ldots c_{K}} = \sum_{u \in U} n^u_{i,j}, n^v_{c_{i} \ldots c_{K}} = \sum_{j \in L} n^v_{i,j} \).

Proof. Appendix. \( \square \)

Once the current location \( i \) is fixed, the first term in the r.h.s of (10) is constant over all users. The second term can be seen as a weighted sum of the empirical transition kernels of all users (i.e., \( n^u_{i,j}/n^u_v, \forall v \in U \) ). The corresponding weight for user \( v \) (the term inside the brackets in (10)) which quantifies how much we account for user-\( v \)’s trajectory in the prediction of user \( u \) at the current location \( i \) is mainly decided by clusters of \( v \) sampled from multiple iterations in CAB and \( v \)’s visiting frequency to the location \( i \) (i.e., \( n^v_{i,j} \)). It is of importance to note that the weight of \( v \) can be seen as a notion of similarity between \( u \) and \( v \), because, as the number of sampled clusters in which both \( u \) and \( v \) are involved increases, the weight of \( v \) in (10) accordingly increases. Also, if \( v \) has relatively high \( n^v_{i,j} \) compared to other users (i.e., \( v \) has accumulated more observations at \( i \) than other users), higher weight is assigned to \( v \).
V. Consistency and Performance Guarantee

In this section, we analyze to what extent $E_g[\theta^u|x^u]$ (that is well approximated, when $B$ is large, by $\hat{\theta}^u$ derived in the CAB algorithm) is close to $E[\theta^u|x^u]$, the expectation under the true prior $\mu$. We are mainly interested in the regime where the user population $\mathcal{U}$ becomes large, while the number of observations $n^u$ for each user remains bounded. This regime is motivated by the fact it is often impractical to gather long trajectories for a given user, while the user population available may on the contrary be very large. For the sake of the analysis, we assume that the length $n^u$ of user-$u$’s observed trajectory is a random variable with distribution $p \in \mathcal{P}(\mathbb{N})$, and that the lengths of trajectories are independent across users. We further assume that the length is upper bounded by $\overline{n}$, e.g., $\overline{n} = \max\{n: p(n) > 0\} < \infty$.

Since the length of trajectories is bounded, we cannot ensure that $|E_g[\theta^u|x^u] - E[\theta^u|x^u]|$ is arbitrarily small. Indeed, for example if users’ trajectories are of length 2 only, we cannot group users into clusters, and in turn, we can only get a precise estimate of the transition kernels averaged over all users. In particular, we cannot hope to estimate $E(\theta^u|x^u)$ for each user $u$. Next we formalize this observation. We denote by $\mathcal{H}_n \subseteq \mathcal{P}^n$ the set of possible trajectories of length less than $\overline{n}$. With finite-length observed trajectories, there are distributions $\nu \in \mathcal{P}(\Theta)$ that cannot be distinguished from the true prior $\mu$ by just observing users’ trajectories, i.e., these distributions induce the same law on the observed trajectories as $\mu$; $P_\nu = \mathbb{P}$ on $\mathcal{H}_n$ (here $P_\nu$ denotes the probability measure induced under $\nu$, and recall that $\mathbb{P}$ is the probability measure induced by $\mu$).

We prove that when the number of observed users grows large, $|E_g[\theta^u|x^u] - E[\theta^u|x^u]|$ is upper-bounded by the performance provided by a distribution $\nu$ indistinguishable from $\mu$, which expresses the consistency of our inference framework. Before we state our result, we introduce the following two related notions:

KL $\epsilon$-neighborhood: the Kullback-Leibler $\epsilon$-neighborhood $K_{\epsilon}(\mu)$ of a distribution $\mu \in \mathcal{P}(\Theta)$ with respect to $\mathcal{H}_n$ is defined as the following set of distributions:

$$K_{\epsilon}(\mu) = \{\nu \in \mathcal{P}(\Theta) : KL(\mu, \nu) < \epsilon\},$$

where $KL(\mu, \nu) = \sum_{x \in \mathcal{H}_n} P_\mu(x) \log \frac{P_\mu(x)}{P_\nu(x)}$.

KL support: The distribution $\mu$ is in the Kullback-Leibler support of a distribution $g \in \mathcal{P}(\mathcal{P}(\Theta))$ with respect to $\mathcal{H}_n$ if $g(K_{\epsilon}(\mu)) > 0$ for all $\epsilon > 0$.

Theorem 2 If $\mu \in \mathcal{P}(\Theta)$ is in the KL-support of $g$ with respect to $\mathcal{H}_n$, then we have, $\mu$-almost surely, for any $i, j \in \mathcal{L}$,

$$\lim_{|U| \to \infty} \frac{E_g[\theta^u_{i,j}|X^u] - E[\theta^u_{i,j}|X^u]}{\sup_{\nu \in \mathcal{P}(\Theta)}\left|E[\theta^u_{i,j}|X^u] - E[\theta^u_{i,j}|X^u]\right|} \leq \frac{\max_{\nu \in \mathcal{P}(\Theta)}\sup_{P_\nu = P_\mu} E[\theta^u_{i,j}|X^u]}{\sup_{\nu \in \mathcal{P}(\Theta)}\left|E[\theta^u_{i,j}|X^u] - E[\theta^u_{i,j}|X^u]\right|}. \quad (11)$$

Proof. Appendix. □

The r.h.s. of (11) captures the performance of an algorithm that would perfectly estimate $E[\theta^u|x^u]$ for the worst distribution $\nu$, which agrees with the true prior $\mu$ on $\mathcal{H}_n$. Note that in our framework, for the prior $g \in \mathcal{P}(\mathcal{P}(\Theta))$, we use is a DP mixture $DP(G_\theta, \alpha)$, with a base measure $G_\theta \in \mathcal{P}(\Theta)$ having full support $\Theta$. Therefore, the KL-support of $g$ is here the whole space $\mathcal{P}(\Theta)$; it thus contains $\mu$, and the theorem applies.

As far as we are aware, Theorem 2 presents the first performance result on inference algorithms using DP mixture models with indirect observations. By indirect observations, we mean that the kernels $\theta^u$ cannot be observed directly, but are revealed only through the trajectories $x^u$. Most existing analysis do not apply in our setting, as these papers aim at identifying conditions on the Bayesian prior $g$ and on the true distribution $\mu$ under which the Bayesian posterior $g(\theta|x)$ will converge (either weakly or in $L_1$-norm) to $\mu$ in the limit of large population size. Hence, existing analysis are concerned with direct observations of the kernels $\theta^u$.

VI. Simulation Experiments

In this section, we explore the performance of our inference algorithm CAB using artificially generated mobility traces according to the model described in III-C.

A. Setup

We randomly generate the trajectories of users over a set $\mathcal{L}$ of $L = 20$ locations by synthetic transition kernels, $\theta^u, \forall u \in \mathcal{U}$. The $\theta^u$’s are independently sampled from a given prior $\mu$, and we consider two scenarios depending on the level of clustering present in $\mu$.

Perfect Clusters. In this case, we assume that there are $C$ perfect clusters of equal average sizes, i.e., $\mu(\theta) = \frac{1}{C} \sum_{c=1}^C \delta_{\bar{\theta}_c}(\theta)$.

Imperfect Clusters. In this case, we set $\mu(\theta) = \frac{1}{C} \sum_{c=1}^C \kappa_{\bar{\theta}_c}(\theta)$, where $\kappa_{\bar{\theta}_c}(\theta)$ is a distribution centered at $\bar{\theta}_c$ and with variance $\epsilon^2 L$.

In both scenarios, the $\bar{\theta}_c$’s are independently drawn uniformly at random from $\Theta = \{\theta : \forall i \in \mathcal{L}, \sum_{j \in \mathcal{L}} \theta_{ij} = 1\}$, and the number $C$ of clusters is fixed equal to 10. To assess the performance of various algorithms, we consider cases where the number of users is 100 and 1000. We compare the performance of the CAB predictor with that of the following predictors: (i) The Markov predictor [4], referred to as Markov in the plots: under this algorithm, the prediction of the next location of user $u$ currently at location $i$ is arg max $\frac{n_{u,i}}{\overline{n}}$, and hence depends on the observed trajectory of user $u$ only. (ii) The spectral clustering-based prediction algorithm (SC-KL) which is a variation of SC-PPK [13]: under this algorithm, users are first clustered using a simple spectral method, and the prediction for a given user leverages the trajectories of users in the same clusters. Note that SC-KL and SC-PPK [13] require the knowledge of the number of clusters, and are less flexible than the CAB predictor.

The parameters in CAB are $B = 8$, $K = 5$, and $M = 30$.

B. Prediction Accuracy

Under the various algorithms, we evaluate the accuracy $AC^u(\hat{\theta}^u)$ of the prediction for user $u$, as defined below.

$$AC^u(\hat{\theta}^u) = \mathbb{P}(X^u_{n+1} = \arg \max_{\hat{\theta}^u_{i,j}} X^u_{n+1,i,j}).$$
Naturally, the highest possible prediction accuracy for user $u$ is $AC^u(\theta^u)$, and we define the (relative) error of a predictor for user $u$ by:

$$error_u = AC^u(\theta^u) - AC^u(\hat{\theta}^u).$$

Fig. 1(a) presents the error, averaged over all users, of each predictor in the perfect cluster scenario. The error is evaluated vs. the observed trajectory lengths (all users are assumed to have trajectories of equal length). As the length of trajectories increases, the error of all predictors decreases, since the longer history trajectory leads to a more precise estimation of $\theta^u$. When we increase the number of users from 100 to 1000, the performance of CAB is significantly improved, which indicates that CAB efficiently exploit the clustered structure of the data, as expected. Note that the performances under CAB and SC-KL are similar, but here in the implementation of SC-KL, the number of clusters is assumed to be known. This indicates that CAB efficiently learn the number of clusters hidden in the data.

In Fig. 1(b), we consider the imperfect cluster scenario. We fix the number of users to 200, and vary $\epsilon$ which measures the gap between this scenario and the perfect cluster scenario. The observed trajectories are of the same length $n$ for all users, fixed to either 50 or 500. We plot the error averaged over all users. As expected, when $\epsilon$ increases, the error under CAB increases. When $\epsilon$ is very low, CAB and SC-KL perform similarly and much better than the Markov predictor. When $\epsilon$ is large (which means that the data does not really exhibit a clustered structure), the performance under SC-KL is very poor and worse than under the Markov predictor. This is expected as SC-KL is not adaptive, and always aims to improve the prediction of user-$u$’s mobility. To this aim, we first define the empirical accuracy of an estimator $\hat{\theta}^u$ of user-$u$’s transition kernel:

$$AC^u(\hat{\theta}^u) = \frac{1}{n^u - 1} \sum_{t=2}^{n^u} \| (a^u_t = \arg \max_j \hat{\theta}^u_{t-1,j}) \|$$

where $\| \cdot \|$ denotes the Jaccard index between two lists $A$ and $B$ is defined as $\frac{A \cap B}{A \cup B}$.  

**Wi-Fi traces** [24]. We use the dataset of [24] where the mobility of 62 users are collected for three months in WiFi networks mainly around Yonsei campus, Korea. The smartphone of each participant periodically scans its radio environment and gets a list of mac addresses of available access points (APs). To map these lists of APs collected over time to a set of locations, we compute the Jaccard index between two lists of APs scanned at different times. If two lists of APs have a Jaccard index higher than 0.5, these two lists are considered to correspond to a same geographical locations [23]. From the constructed set of locations, we then construct the trajectories of the various users.

**ISP traces** [25]. We also use the dataset provided by Orange where the mobility of 50000 subscribers in Senegal are measured over two weeks. More specifically, we use the SET2 data [25], where the mobility of a given user is reported as a sequence of base station (BS) ids, and time stamps.

In each dataset, we first restrict our attention to a subset $\mathcal{L}$ of frequently visited locations. We select the 116 and 80 most visited locations in Wi-Fi traces and ISP traces datasets, respectively. The most of locations in sets $\mathcal{L}$ are presented in Fig 2. We then re-construct users’ trajectories by removing locations not in $\mathcal{L}$. For the ISP dataset, we extract 200 users (randomly chosen among users who visited at least 10 of the locations in $\mathcal{L}$). From the re-constructed trajectories, we observe a total number of transitions from one location to another equal to 8194 and 13453 for the WiFi and ISP dataset, respectively. The CDF of the length of trajectories of users is plotted in Fig 3(a).

**Users’ similarity.** Before actually evaluating the performance of various prediction algorithms, we wished to assess whether users exhibit similar mobility patterns, that could in turn be exploited in our predictions. Here, we just test the similarity of pairs of users only. More precisely, we wish to know whether the observed trajectory of user $u$ could be aggregated to that of user $v$ to improve the prediction of user-$u$’s mobility. To this aim, we first define the empirical accuracy of an estimator $\hat{\theta}^u$ of user-$u$’s transition kernel:

$$AC^u(\hat{\theta}^u) = \frac{1}{n^u - 1} \sum_{t=2}^{n^u} \| (a^u_t = \arg \max_j \hat{\theta}^u_{t-1,j}) \|$$

where $\| \cdot \|$ denotes the Jaccard index between two lists $A$ and $B$ is defined as $\frac{A \cap B}{A \cup B}$.  

---

**VII. Evaluation on Mobility Dataset**

**A. Mobility Traces**

We evaluate the performance of CAB using two sets of mobility traces collected on a WiFi and cellular network,
We hence define the arg max and there are 173 (out of 200) such users. These numbers are high, we found 19 (out of 62) such users, whereas in the ISP traces. We observe that 1.65% and 5% of user pairs have similarity higher than 0.5 for the WiFi and ISP traces. We also ISP traces.

Let $\hat{\theta}^{uv}$ be the maximum likelihood estimator of $\theta^u$ given $x^u$ (i.e., $\hat{\theta}_{i,j}^{uv} = \frac{n^u_{ij}}{n^u}$, $\forall i,j \in \mathcal{L}$). Intuitively, user-v’s trajectory is useful to predict the mobility of user u if $\hat{\theta}^{uv}$ has a high empirical accuracy for user u, i.e., if $AC^u(\hat{\theta}^{uv})$ is high. We hence define the similarity $sim(u,v)$ of users $u$ and $v$ as $sim(u,v) = AC^v(\hat{\theta}^{uv})/AC^u(\hat{\theta}^{uu})$. Note that the notion of similarity is not symmetric (in general $sim(u,v) \neq sim(v,u)$), and it always takes its value between 0 and 1.

Fig 4(b) plots the CCDF of the similarity between all possible pairs of 62 users in Wi-Fi traces and 200 users in ISP traces. We observe that 1.65% and 5% of user pairs have similarity higher than 0.5 for the Wifi and ISP traces. We also computed the number of users having at least one user with whom the similarity is higher than 0.5. In the Wi-Fi traces, we found 19 (out of 62) such users, whereas in the ISP traces there are 173 (out of 200) such users. These numbers are high, and motivate the design of cluster-aided predictors.

Fig. 4(a) and (b) present the similarity between 62 users in Wi-Fi trace and 100 users in ISP subscriber dataset. To provide meaningful plots, we have ordered users so that pairs of users with high similarity are neighbours (to this aim, we have run the spectral clustering algorithm [13] and re-grouped users in the identified clusters). From these plots, the similarity of users is apparent, however we also clearly observe that perfect clusters (as considered in [17]) do not really exist.

B. Prediction Accuracy

1) Tested Predictors: We assess the performance of six types of predictors: the order-1 Markov predictor (Markov [4]), the order-2 Markov predictor with fallback (Markov-O(2) [4]), AGG, CAB and CAB$^C$, AGG$^C$. Before describing each predictor, we briefly introduce some notions of training data available at a given time. The time stamp of the arrival at $t$-th location on user-u’s trajectory is denoted by $d^u_t \in \mathbb{R}$, and $n^u(d)$ is the length of trajectory collected by $u$ before time $d$ (i.e., $n^u(d) = \max\{s|d^u_s < d\}$). The collection of users’ trajectories available for a prediction at time $d$ is denoted by $\mathcal{X}^d$, $\mathcal{X}^d = \{x^v_d : v \in \mathcal{U}, \text{where } x^v_d = (x^v_1, ..., x^v_{n^v_d})\}$. The prediction for $x^v_t$ is denoted by $\hat{x}^v_t$. In order to predict $\hat{x}^v_t$, the Markov, AGG and CAB predictors first estimate $\hat{\theta}^{uv}$ based on $\mathcal{X}^d$ (this estimate is denoted by $\hat{\theta}^{uv}$) in different ways, and predicts $\hat{x}^v_t$ accordingly. The AGG predictor uses the trajectories of all users to estimate the transition kernel of a given user $u$ (as if there were only a single cluster). The description of the tested order-1 predictors is summarized in Table 1.

CAB$^C$ differs from CAB in that the prediction at time $d$ under CAB$^C$ for user $u$ uses other users’ complete trajectories (i.e., $\mathcal{X}^d(u)$). This corresponds to a case where user $u$ starts moving along her trajectory after other users have gathered sufficiently long trajectories. AGG$^C$ operates in a same manner. Markov-O(2) assumes that users’ trajectories are order-2 markov chains, and for the locations where the corresponding order-2 transitions are not observed, Markov-O(2) falls back to the Markov predictor.

The parameters $B$, $K$ and $M$ for CAB and CAB$^C$ are set to 8, 3 and 30, respectively.

2) Results: We assess the performance of the various algorithms using two main types of metrics. The first metric, referred to as the Cumulative Accurate Prediction Ratio (CAPR), is defined as the fraction of accurate predictions for all users up to time $d$:

$$CAPR_{time} = \frac{1}{\sum_{u \in \mathcal{U}}(n^u(d)-1)\sum_{s=2}^{n^u(d)}\sum_{v \in \mathcal{U}}\mathbb{I}(\hat{x}^v_s = x^v_s)}.$$
We also introduce a similar metric that captures the cumulative accuracy of predictions after observing \( t \) different locations on users’ trajectories:

\[
CAPR = \frac{1}{(t-1)} \sum_{u \in U} \sum_{n \geq t} \sum_{s=2}^{t} \mathbb{1}(x^n_s = \hat{x}_n^u).
\]

The second type of metrics concerns the instantaneous accuracy of the predictions. The Instantaneous Accurate Prediction Ratio (IAPR) after observing \( t \) different locations on users’ trajectories is defined as follows.

\[
IAPR = \frac{1}{\sum_{u \in U} \sum_{n \geq t} \sum_{s=2}^{t} n^u_{x^n_{s-1}} \hat{x}^u_{x^n_{s}}}. 
\]

IAPR is the fraction of accurate predictions made after observing trajectories of length \( t \).

Fig. 5(a)-(b) present \( CAPR_{time} \) as a function of time \( d \) for various algorithms and for the two mobility traces. CAB outperforms all other algorithms at any time. The improvement over Markov and Markov-O(2) can be as high as 65%. This illustrates again the performance gain that can be achieved when exploiting users’ similarities. Note Markov-O(2) does not outperform Markov, which was also observed in [6]. In the following, we only evaluate the performance of the Markov predictor, and do not report that of its order-2 equivalent.

In Fig. 5(c)-(f), we plot the CAPR and IAPR as a function of the length \( t \) of the observed trajectory. In Fig. 5(c) and (d), when the collected trajectory is not sufficient (i.e., \( t = 10 \)), CAB\(^C\) and CAB outperform Markov by 64% and 40%, respectively. Regarding the IAPR in Wi-Fi traces, Fig. 5(e) shows that CAB and CAB\(^C\) provide much better predictions than Markov, when the length of trajectory is less than 140. After a sufficient training data is collected, they yield comparable IAPR. In Fig. 5(f), for the ISP traces, the IAPR under CAB and Markov are similar sooner, for trajectories of length greater than 20 only.

In Fig. 5(g) and (h), we evaluate the CAPR and IAPR averaged only over users having at least one user with whom the similarity is higher than 0.5 (see VII-A). These users are referred to as Mobility Friendly (MF) users. In Fig. 5(g), we observe that for MF users, the gain of CAB\(^C\) and CAB becomes really significant, i.e., when \( t=10 \), the CAPR of CAB\(^C\) and CAB outperform that of Markov by 102% and 65%, respectively. Also note that CAB\(^C\) becomes significantly better than CAB for MF users. This is explained by the fact that we can predict the mobility of MF users much more accurately if we have a long history of the mobility of users they are similar to. The performance for MF users in the ISP traces is not presented, because there, most of users (i.e., 86%) are already MF users.

3) Exploiting Similarities in CAB: Recall that, by the weight of the empirical transition kernel of user \( v \) in (10), we can quantify to what extent the observed trajectory of user \( v \) is taken into account in the estimate of \( \hat{\theta}^u \). When summing the weight of user \( v \) over all locations \( i \), we get an aggregate indicator capturing how \( v \) impacts the prediction for user-\( u \)’s mobility. To understand how many users actually impact the prediction for user \( u \) in the CAB predictor, we may look at the cardinality of the set of users whose aggregate indicator exceeds a given threshold:

\[
\left\{ \nu \left| \sum_{i \in L} \sum_{c_1, \ldots, c_K} \xi_{c_1, \ldots, c_K} n_i^{c_k} \sum_{k=1}^{K} \mathbb{1}(v \in c_k) \left| \mathcal{U} \right| \sum_{k=1}^{K} n_i^{c_k} \prod_{k=1}^{K} n_i^{c_k} \right| > \frac{1}{\left| \mathcal{U} \right|} \right\},
\]

where \( z \) is a normalization constant to make the sum of aggregate indicators over all users equal to 1. The above set
CAB

is called the set of $u$-similar users.

In Fig. 6, we plot the size of the set of $u$-similar users, averaged over all users $u$, and as a function of the length of trajectories (in days $d$). In case of CAB, the first day, the average sizes are 7 and 110 in Wi-Fi traces and ISP traces, which means that CAB aggressively uses the trajectories of all users for its prediction. When the length of the trajectories increase, the average size decreases to 1.55 after one month in Wi-Fi traces and 2.24 after two weeks in ISP traces. In other words, as data is accumulated, CAB does not use the trajectories of a lot of users for its prediction. In the case of $CAB^C$, we observe a faster decrease of the average size when time passes, which means that $CAB^C$ tend to utilize other users’ data more selectively, even at the beginning. This explains why $CAB^C$ performs better than CAB as observed in Fig. 5.

C. Required Training Data in Dataset

To conclude, we measure the duration of the observed trajectories required to achieve a certain level of performance under various prediction algorithms. This duration is equal to $D = \min \{ d | \frac{\sum_{\theta} \theta^u \cdot d}{\sum_{\theta} \theta^u} \geq \alpha \}$. The definition of empirical accuracy of user $u$ $AC^C(\cdot)$ is presented in [12].

Fig. 7 plots the CDF of $D$ of MF users for three levels (i.e., $th=0.7, 0.9, 0.95$). As expected $CAB^C$ provides the required level performance with much less data than the other algorithms. In other words, using cluster-aided mobility prediction algorithms, we may significantly reduce the required training period.

VIII. CONCLUDING REMARKS

In this paper, we have presented a cluster-aided inference method to predict the mobility of users in wireless networks. This method significantly departs from existing prediction techniques, as it aims at exploiting similarities in the mobility patterns of the various users to improve the prediction accuracy. The proposed algorithm, CAB, relies on Bayesian non-parametric estimation tools, and is robust and adaptive in the sense that it exploits users’ mobility similarities only if the latter really exist. We have shown that our Bayesian prediction framework is asymptotically optimal when the user population grows large, and have presented extensive numerical experiments indicating that CAB outperforms any other existing prediction algorithms. Many interesting questions remain about the design of CAB. In particular, we plan to investigate how to set its parameters ($B, K,$ and $M$) to achieve an appropriate trade-off between accuracy and complexity. These parameters could also be modified in an online manner while the algorithm is running to adapt to the nature of the data. We further plan to apply the techniques developed in this paper to various kind of mobility, e.g., we could investigate how users dynamically browse the web, and use our framework to predict the next visited webpage.

REFERENCES

[1] A. J. Nicholson and B. D. Noble, “Breadcrumbs: forecasting mobile connectivity,” in Proceedings of ACM MobiCom, 2008.
[2] V. A. Siris and D. Kalyvas, “Enhancing mobile data offloading with mobility prediction and prefetching,” ACM SIGMOBILE Mobile Computing and Communications Review, vol. 17, no. 1, pp. 22–29, 2013.
[3] Y. Chon, E. Talipov, H. Shin, and H. Cha, “Smartdc: Mobility prediction-based adaptive duty cycling for everyday location monitoring,” IEEE Transactions on Mobile Computing, no. 3, pp. 512–525, 2014.
[4] L. Song, D. Kotz, R. Jain, and X. He, “Evaluating next-cell predictors with extensive wi-fi mobility data,” Mobile Computing, IEEE Transactions on, vol. 5, no. 12, pp. 1633–1649, 2006.
[5] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell, “Nextplace: a spatio-temporal prediction framework for pervasive systems,” in Pervasive Computing. Springer, 2011, pp. 152–169.
[6] Y. Chon, H. Shin, E. Talipov, and H. Cha, “Evaluating mobility models for temporal prediction with high-granularity mobility data,” in Pervasive Computing and Communications (PerCom). IEEE, 2012.
[7] M. Feder, N. Merhav, and M. Gutman, “Universal prediction of individual sequences,” Information Theory, IEEE Transactions on, vol. 38, no. 4, pp. 1258–1270, 1992.
[8] P. Jacquet, W. Szpankowski, and I. Apostol, “A universal predictor based on pattern matching,” Information Theory, IEEE Transactions on, vol. 48, no. 6, pp. 1462–1472, 2002.
[9] T. S. Ferguson, “A bayesian analysis of some nonparametric problems,” The annals of statistics, pp. 209–230, 1973.
[10] J. D. McAuliffe, D. M. Blei, and M. I. Jordan, “Nonparametric empirical bayes for the dirichlet process mixture model,” Statistics and Computing, vol. 16, no. 1, pp. 5–14, 2006.
[11] N. Merhav, M. Feder, and M. Gutman, “Some properties of sequential predictors for binary markov sources,” Information Theory, IEEE Transactions on, vol. 39, no. 3, pp. 887–892, 1993.
[12] P. Smyth et al., “Clustering sequences with hidden markov models,” Advances in neural information processing systems, pp. 648–654, 1997.
[13] T. Jebara, Y. Song, and K. Thadani, “Spectral clustering and embedding with hidden markov models,” in Machine Learning: ECML 2007. Springer, 2007, pp. 164–175.
[14] O. Kallenberg, Foundations of modern probability. springer, 2002.
[15] D. Blackwell and J. B. MacQueen, “Ferguson distributions via pólya urn schemes,” The annals of statistics, pp. 353–355, 1973.
APPENDIX

A. Proof of Lemma 7

Observe that in view of (9), we have:

$$G_{0}^{k+1}(d\theta) = \sum_{c} \omega_{c}^{k} P_{0}(x^{c}) G_{0}^{k}(d\theta),$$  \hspace{1cm} (13)

where the sum is over all possible partitions of the set of users \( U \) in clusters and the weight \( \omega_{c}^{k} \) is

$$\omega_{c}^{k} = \frac{n_{c}^{k}}{B|U| \int_{\theta} P_{0}(x^{c}) G_{0}^{k}(d\theta)},$$  \hspace{1cm} (14)

with \( n_{c}^{k} = \sum_{b=1}^{B} \sum_{k \in U} I(\theta^{u,b,k} = c) \). Recursively replacing \( G_{0}^{k} \) in (13) with \( G_{0}^{k-1} \) and putting \( G_{0}^{0} = \text{Uniform}(\Theta) \), we obtain another expression of \( G_{0}^{K} \) as

$$G_{0}^{K}(d\theta) = \sum_{c_{1},..c_{K-1}} \prod_{k=1}^{K-1} \omega_{c_{k}}^{k} P_{0}(x^{c_{k}}) d\theta,$$  \hspace{1cm} (15)

where the sum \( \sum_{c_{1},..c_{K-1}} \) is \( \sum_{c_{1} \in C_{1}} \cdots \sum_{c_{K-1} \in C_{K-1}} \) where \( C_{k} \) is a set of every cluster sampled at \( k \)-th iterations, i.e.,

$$C_{k} = \{c \} \sum_{b=1}^{B} \sum_{u \in U} I(\theta^{u,b,k} = c) > 0 \}.$$  \hspace{1cm} (19)

We can further obtain the recursive expression of the weights \( \omega_{c}^{k} \) by plugging (15) into (14):

$$\omega_{c}^{K} = \frac{n_{c}^{K}}{B|U| \sum_{c_{1},..c_{K-1}} \prod_{k=1}^{K-1} \omega_{c_{k}}^{k}},$$  \hspace{1cm} (16)

$$\xi_{c_{1},..c_{K}} = \int_{\theta} \prod_{k=1}^{K} P_{0}(x^{c_{k}}) d\theta$$  \hspace{1cm} (17)

$$= \prod_{i \in \mathcal{E}} \Gamma(1 + \sum_{k=1}^{K} n_{i,j}^{c_{k}})$$  \hspace{1cm} (18)

where

$$n_{c}^{u} = \sum_{u \in \mathcal{E}} n_{i,j}^{u} \quad n_{c}^{u} = \sum_{j \in \mathcal{L}} n_{i,j}^{c_{k}}.$$

Then, using equations (9), (15) and (17), we get an expression for \( \theta_{i,j}^{u,k} \): In (9), replacing the denominator of each sample \( b \) with \( \omega_{c_{b},b,K} \), and plugging (15) into numerator, we arrive at

$$\hat{\theta}_{i,j}^{u,k} = \frac{1}{B} \sum_{b=1}^{B} \frac{\omega_{c_{b},b,K} B|U|}{n_{c_{b},b,K}} \sum_{c_{1},..c_{K-1}} \int_{\theta} P_{0}(x^{c_{b}..c_{K}})$$

$$\quad \prod_{k=1}^{K-1} P_{0}(x^{c_{k}}) \omega_{c_{k}}^{k} d\theta$$

$$= \sum_{c_{1},..c_{K}} \xi_{c_{1},..c_{K}} \prod_{i \in \mathcal{E}} \Gamma(1 + \sum_{k=1}^{K} n_{i,j}^{c_{k}}) \prod_{k=1}^{K} \omega_{c_{k}}^{k}$$  \hspace{1cm} (19)

Rearranging (19), we arrive at (10).

B. Proof of Theorem 2

The proof of Theorem 2 relies on the following two lemmas.

**Lemma 3**: If \( \mu \in \mathcal{P}(\Theta) \) is in the KL-support of \( g \) with respect to \( \mathcal{H}(\Theta) \), then \( g(K, \mu(\phi), \lambda^{(K)}) \rightarrow 1 \) for all \( \epsilon > 0 \), \( \mu \)-almost surely.
The above lemma is a perfect analog of a similar statement for Bayesian consistency with direct observations (see [23], Theorem 6.1 and its corollary). The proof also goes through essentially in the same way; therefore, we do not provide it here. This first lemma states that the set of distributions $\nu$ that do not agree with the true prior $\mu$ on $\mathcal{H}_\pi$ according to the KL distance $KL_{\pi}(\mu, \nu)$ w.r.t. $\mathcal{H}_\pi$ has a vanishing mass under the posterior distribution $g|X^u$, $\mu$-a.s. However, this does not guarantee that the set of distributions $\nu$ with $0 < KL_{\pi}(\mu, \nu) \leq \epsilon$ will have a negligible impact on the estimates $E_\nu[\theta^n|X^u]$. Indeed, for this we need continuity with respect to the KL distance over $\mathcal{H}_\pi$, which the next lemma provides.

**Lemma 4** Under the assumptions of Lemma [23] for any bounded continuous $f : \Theta \to \mathbb{R}$, $\lim_{\epsilon \to 0} \sup_{\nu \in \mathcal{P}(\Theta)} |E_\nu[f] - \mathbb{E}[f]| = \sup_{\nu \in \mathcal{P}(\Theta)} |E_\nu[f] - \mathbb{E}[f]|$.

*Proof.* Let $\rho$ be the metric on $\Theta$. We use the associated Wasserstein metric $d_\rho$ on $\mathcal{P}(\Theta)$:

$$d_\rho(\mu, \nu) = \inf_{\pi_1 = \mu, \pi_2 = \nu} \left\{ \int \rho(\theta, \lambda) \pi(d\theta, d\lambda) \right\},$$

where $\pi_1$ and $\pi_2$ are the first and second marginals of $\pi$, respectively. It is well-known (see [26]) that the space $(\mathcal{P}(\Theta), d_\rho)$ is compact, complete and separable, as $(\Theta, \rho)$.

Let $\delta > 0$ and let $(\epsilon_k) \subseteq \mathbb{R}_0^+$ be a sequence converging to 0. For all $k \in \mathbb{N}$, let $\nu_k \in \mathcal{P}(\Theta)$ such that $|E_{\nu_k}[f] - \mathbb{E}[f]| \geq \sup_{K_{L_{\pi}(\mu, \nu)} \leq \epsilon_k} |E_\nu[f] - \mathbb{E}[f]| - \delta$.

By compactness of $(\mathcal{P}(\Theta), d_\rho)$, there exists a converging subsequence $(\nu_{k_i})$ of $(\nu_k)$, and a corresponding subsequence $(\epsilon_{k_i})$ of $(\epsilon_k)$; let us call $\nu_{\infty} \in \mathcal{P}(\Theta)$ its limit. Clearly, we have $D_{\infty}(\mu, \nu_\infty) = 0$. Because the Wasserstein distance metricizes weak convergence (see Theorem 6.9 in [26]) and $f$ is bounded and continuous, we have that $\lim_{k \to \infty} E_{\nu_k}[f] = E_{\nu_\infty}[f]$. Thus,

$$\lim_{k \to \infty} |E_{\nu_k}[f] - \mathbb{E}[f]| \geq \lim_{k \to \infty} |E_{\nu_\infty}[f] - \mathbb{E}[f]| - \delta$$

where the last inequality is because the sequence is decreasing. Letting $\delta \to 0$ completes the proof. The opposite inequality is obvious by the definition of $K_{L_{\pi}(\mu, \nu)}$.

*Proof of Theorem [23]* For any bounded continuous $f : \Theta \to \mathbb{R}$, we have

$$|E_\nu[f(\theta^n)|X^u] - \mathbb{E}[f(\theta^n)|X^u]| \leq \|f\|_{\infty} \mathcal{L}(K_{\pi}(\mu)|X^u)$$

According to Lemma [3] the first term in the r.h.s. goes to 0 as $|U| \to \infty$, $\mu$-a.s. The second term can always be upper-