Predictive Distance-Based Mobility Management for Multidimensional PCS Networks

Ben Liang, *Member, IEEE,* and Zygmunt J. Haas, *Senior Member, IEEE*

Abstract—This paper presents a mobile tracking scheme that exploits the predictability of user mobility patterns in wireless PCS networks. In this scheme, a mobile's future location is predicted by the network, based on the information gathered from the mobile's recent report of location and velocity. When a call is made, the network pages the destination mobile around the predicted location. A mobile makes the same location prediction as the network does; it inspects its own location periodically and reports the new location when the distance between the predicted and the actual locations exceeds a threshold. To more realistically represent the various degrees of velocity correlation in time, a Gauss–Markov mobility model is used. For practical systems where the mobility pattern varies over time, we propose a dynamic Gauss–Markov parameter estimator that provides the mobility parameters to the prediction algorithm.

Based on the Gauss–Markov model, we describe an analytical framework to evaluate the cost of mobility management for the proposed scheme. We also present an approximation method that reduces the computational complexity of the cost evaluation for multidimensional systems. We then compare the cost of predictive mobility management against that of the regular, nonpredictive distance-based scheme, for both the case with ideal Gauss–Markov mobility pattern and the case with time-varying mobility pattern.

The performance advantage of the proposed scheme is demonstrated under various mobility patterns, call patterns, location inspection cost, location updating cost, mobile paging cost, and frequencies of mobile location inspections. As a point of reference, prediction can reduce the mobility management cost by more than 50% for all systems, where a the mobile users have moderate mean velocity and where performing a single location update is as least as expensive as paging a mobile in one cell.

Index Terms—Distance-based location management, dynamic parameter estimation, fluid flow, Gauss–Markov model, mobility pattern, predictive mobility management, random walk, wireless networking.

I. INTRODUCTION

IN THE operation of wireless personal communication service (PCS) networks, mobility management deals with the tracking, storage, maintenance, and retrieval of mobile location information. Two commonly used standards, the EIA/TIA Interim Standard 41 in North America [1] and the Global System for Mobile Communications in Europe [22] partition their coverage areas into a number of location areas (LAs), each con-

sisting of a group of cells. When a mobile enters an LA, it reports to the network the information about its current new location (*location update*). When an incoming call arrives, the network simultaneously pages the mobile (*terminal paging*) in all cells within the LA, where the mobile currently resides. In these standards, the LA coverage is fixed for all users and cannot adapt to different and differing user traffic and mobility patterns.

Dynamic LA management schemes have been proposed in [15], [18], and [32], where an LA is defined per mobile, as to adapt to its traffic and mobility patterns. However, when a call arrives, the system still needs to page the mobile in all cells within the LA, an operation that is costly in most cases.

Dynamic mobility management schemes [2], [3], [5], [8], [14], [19], [21] discard the notion of LA borders altogether. A mobile in these schemes updates its location[1] based on either elapsed time, number of crossed cell borders, or traveled distance. All these parameters can be dynamically adapted to each mobile's traffic and mobility patterns, hence, providing better cost effectiveness than the LA scheme. In schemes that do not rely on LAs, upon call arrival, the network pages the destination mobile using a selective paging scheme [24], starting from the cell location where the mobile last updated and outwards, in a shortest-distance-first order. With the assumption of a random-walk mobility model, this paging scheme is equivalent to the highest-probability-first algorithm, which, indeed, incurs the minimum paging cost. In particular, in the distance-based scheme, a mobile performs location update whenever it is some threshold distance away from the location where it last updated. For a system with the memoryless random-walk mobility pattern, the distance-based scheme has been proven to result in less location updating and paging costs than schemes based on time or number of cell boundary crossings [5].

However, in systems where a user's future velocity is correlated with its past and current velocity (i.e., nonmemoryless), the highest probability location of a mobile is generally not the cell where the mobile last reported. Thus, predicting the mobile's current location when a call arrives would result in less paging traffic. Consequently, a mobility management scheme that takes advantage of the predictability of the mobiles' location can lower the mobility management cost.

In our proposed predictive distance-based mobility management scheme, the future location of a mobile is predicted based

[1]We do not address in this paper the exact mechanism by which a mobile monitors its location and velocity. A mobile may determine its location through a variety of methods, including the Global Positioning System, signal triangulation, base-station self-identifying beacons, or a combination of the above. It may then average position displacement over time to find its velocity. Other methods and related references on mobile location and velocity determination can be found in [13].

on the probability density function (PDF) of the mobile's location, which is given by a Gauss–Markov mobility model based on its location and velocity at the time of the recent location updates. The prediction information is made available to both the network and the mobiles. Therefore, a mobile is aware of the network's prediction of its location. The mobile checks its position periodically (*location inspection*) and performs location update whenever it reaches some threshold distance (*updating distance*) away from the predicted location. To locate a mobile, the network pages the mobile starting from the predicted location and outwards, in a shortest-distance-first order, until the mobile is found.

Besides the application of prediction, another feature that sets our work apart from the original investigation of distance-based mobility management is the notion of sporadic location inspections. In a distance-based updating scheme, a mobile needs to monitor its location and/or velocity. In practical systems, this cannot be done continuously due to bandwidth and computation power limitations. We assume that a mobile measures its location and velocity periodically, at the location inspection instants. We further study how the optimal location inspection period is affected by the mean, variance, and memory level of a mobile's velocity. Previous work on distance-based mobility management schemes [2], [5], [14], [21] do not incorporate this assumption.

Summaries of many aforementioned mobility management schemes can be found in the survey papers [4] and [30]. A preliminary version of the predictive distance-based mobility management scheme, based on a one-dimensional (1-D) system and without considering the cost of location inspections, is presented in [20] and summarized in [30]. Other related work include the combination of timer, movement, and distance approaches into a general state-based framework in [26] and an information-theoretic approach in [6]. Most recently, a cell-shape independent stochastic model has been proposed in [31], which provides an iterative algorithm to compute the optimal updating threshold for distance-based mobility management.

The rest of this paper is organized as follows. In Section II, we describe the Gauss–Markov mobility model and the prediction algorithm. We demonstrate the broad applicability of the Gauss–Markov model, derive the optimal location prediction, and introduce an algorithm to dynamically estimate the Gauss–Markov parameters when the mobile movement pattern is time varying. Section III presents the analytical framework for evaluating the cost of the predictive distance-based scheme. This cost is defined as the sum of a mobile's location inspection cost, location updating cost, and the cost incurred in paging the mobile. The numerical results are presented in Section IV. We study how the optimal location inspection period and updating distance are affected by the traffic pattern, the mobility pattern, and the relative costs of location inspection, location updating, and paging. To evaluate the performance gain of the predictive scheme, the cost of the nonpredictive distance-based scheme is measured through simulations, and is compared with the predictive scheme. Our results indicate that, in general, prediction can significantly lower the mobility management cost. We show that the proposed scheme performs well, even when used in conjunction with a suboptimal parameter estimation al-gorithm, in systems with time-varying mobility pattern. Furthermore, we explain why the performance gain of the predictive scheme does not necessarily increase as the memory level of a mobile's movement increases. Finally, concluding remarks are given in Section V.

II. SYSTEM DESCRIPTION

A. Gauss–Markov Mobility Model

A mobile user usually travels with a destination in mind. Furthermore, the change of a mobile's velocity within a short time is limited due to physical restrictions. Therefore, a mobile user's future location and velocity are likely to be correlated with its past and current location and velocity. Thus, the memoryless nature of the random-walk model makes it unsuitable to represent such behavior. Another widely used mobility model in cellular network analysis is the fluid-flow model [32], [33]. The fluid-flow model is suitable for vehicle traffic on highways but for not pedestrian movements with frequent "stop-and-go" behavior. A discrete Markovian model is reported in [5]. However, in this model, the velocity of the mobiles is overly simplified and characterized by three states only. A memoryless Brownian motion with drift model is used in [25] and [8]. It is an improvement to the random-walk model, but it still cannot represent the time correlation in a mobile's velocity. In this paper, we introduce a mobility model based on the Gauss–Markov process [11], [23], [29]. The Gauss–Markov process has appeared in diverse fields, such as the theory and applications of signal estimation and economic forecast. We postulate that it can be used to model the mobile movement in a PCS network, since it captures the essence of the correlation of a mobile's velocity in time. Furthermore, the Gauss–Markov model represents a wide range of user mobility patterns, including, as its two extreme cases, the random-walk and the constant velocity fluid-flow models.

1) 1-D Case: For simplicity, we first illustrate the Gauss–Markov mobility model in a 1-D system. In this model, a mobile's velocity is assumed to be correlated in time and modeled by a Gauss–Markov random process. In continuous time, a stationary Gaussian process $v(t)$ is a Gauss–Markov process if it has the autocorrelation function [11], [23], [29]

$$R_v(\tau) = \mathrm{E}[v(t)v(t+\tau)] = \sigma^2 e^{-\beta|\tau|} + \mu^2 \qquad (1)$$

where σ^2 is the variance of $v(t)$, μ is its mean, and $\beta \geq 0$ determines the degree of memory. When $\mu = 0$, (1) is also sometimes called the Ornstein–Uhlenbeck solution of the Brownian motion with zero restoring force [23].

We define a discrete version of the mobile velocity with $v_n = v(n\Delta t)$ and the memory level $\alpha = e^{-\beta \Delta t}$, where Δt is normalized to unity throughout this paper.[2] Then the discrete Gauss–Markov process described by (1) can be represented by the following recursive realization:

$$v_n = \alpha v_{n-1} + (1-\alpha)\mu + \sigma \sqrt{1-\alpha^2} w_{n-1} \qquad (2)$$

[2]In a practical system design, the value of Δt should match the parameters of each individual system. It should not be too small, such that the model identification and system optimization becomes intractably complex. It also should not be too large, such that the details of a user's movement is lost. However, an investigation into the optimal selection of Δt is outside the scope of this paper.

where $\{w_n\}$ is an uncorrelated Gaussian process with zero mean and unit variance and is independent of $\{v_n\}$.

If the initial velocity v_0 is Gaussian with mean μ and variance σ^2, from (2), we clearly have $R_v(k) = \sigma^2 e^{-\beta k} + \mu^2$. Therefore, (2) satisfies the definition (1). However, since there is no guarantee that v_0 has mean μ and variance σ^2, or even that v_0 is Gaussian, the process described by (2) is generally not stationary. In this case, μ is the asymptotic mean of v_n, and σ is the asymptotic standard deviation of v_n, when n approaches infinity.

Clearly, as α approaches zero, or β approaches infinity, (2) represents a drifting random-walk mobility pattern with mean μ and standard deviation σ, and that as α approaches one or β approaches zero, (2) represents a constant velocity fluid-flow mobility pattern with $v_n = v_0$ for all n. Therefore, not only does the Gauss–Markov model represent a wide spectrum of mobility patterns with various degree of memory, it also includes the random-walk and fluid-flow models as its two extreme cases.

2) Multidimensional Case: It is straightforward to extend the Gauss–Markov mobility model to the two- or three-dimensional systems by using vectors of Gauss–Markov processes. In what follows, we use a two-dimensional (2-D) system to illustrate how this is done.

In the 2-D case, the location and velocity of a mobile is represented by the random vectors $\mathbf{s}_n = [s_n^x, s_n^y]^T$ and $\mathbf{v}_n = [v_n^x, v_n^y]^T$, respectively, where the superscripts denote the dimensions. Similarly, we define $\overline{\alpha} = [\alpha^x, \alpha^y]^T, \overline{\mu} = [\mu^x, \mu^y]^T$, and $\overline{\sigma} = [\sigma^x, \sigma^y]^T$. Then, the 2-D velocity process can be expressed as follows:

$$\mathbf{v}_n = \overline{\alpha} \odot \mathbf{v}_{n-1} + (1 - \overline{\alpha}) \odot \overline{\mu} + \overline{\sigma} \odot \sqrt{1 - \overline{\alpha}^2} \odot \mathbf{w}_{n-1} \quad (3)$$

where $\odot$ denotes element-by-element multiplication, and $\{\mathbf{w}_n\} = [\{w_n^x\}, \{w_n^y\}]^T$ is a 2-D uncorrelated Gaussian process with zero mean and unit variance and is independent of $\{\mathbf{v}_n\}$. In (3), we have made an important assumption that the velocity processes in both dimensions are uncorrelated. It is easy to rewrite (3) to represent the more general case of correlated dimensions. However, as will become obvious in Section III, the cost analysis with even the simpler, uncorrelated model can be highly computationally intensive. The additional dimension of conditional probabilities in the correlated model would render the model practically intractable.

For simplicity of presentation, one may further assume that the velocity has the same memory level α, the same asymptotic mean μ, and the same asymptotic standard deviation σ in both dimensions. In this isotropic case, (3) becomes

$$\mathbf{v}_n = \alpha \mathbf{v}_{n-1} + (1 - \alpha)\mu + \sigma \sqrt{1 - \alpha^2}\, \mathbf{w}_{n-1}. \quad (4)$$

3) Duplicating Other Mobility Models: To our knowledge, no real data are available in the public domain that describe the user mobility pattern in the desired granularity to validate a mobility model. For example, mobile traces reported in [27] and [28] only give coarse movement patterns between large areas, in hourly to daily time frames, aggregated among many users. Therefore, none of the commonly used mobility models in literature can be validated against real user data.

Nevertheless, the wide applicability of the proposed Gauss–Markov model can still be demonstrated. We illustrate how the parameter set (α, μ, σ) can be tuned such that the Gauss–Markov model can duplicate the mobile movements described by three of the most popular mobility models: the random-walk model [25], [8], the fluid-flow model [32], and the random way-point model [7], [9].[3] By showing this, we demonstrate that the Gauss–Markov model is, indeed, a mobility model that can represent a variety of mobility models.

i) The random-walk (RW) model (with drift) has no memory. It can be represented by

$$\mathbf{v}_n^{(\mathrm{RW})} = \overline{\mu}^{(\mathrm{RW})} + \overline{\sigma}^{(\mathrm{RW})} \odot \mathbf{w}_n^{(\mathrm{RW})} \quad (5)$$

where $\overline{\mu}^{(\mathrm{RW})}$ is a vector of drift velocities, $\overline{\sigma}^{(\mathrm{RW})}$ is a vector of velocity standard deviations, and $\{\mathbf{w}_n^{(\mathrm{RW})}\}$ is a vector of uncorrelated Gaussian processes with zero mean and unit variance. Comparing this with (3), the corresponding subfamily of the Gauss–Markov model has the parameters $\{\overline{\alpha} = \mathbf{0}, \overline{\mu} = \overline{\mu}^{(\mathrm{RW})}, \overline{\sigma} = \overline{\sigma}^{(\mathrm{RW})}\}$.

ii) The fluid-flow (FF) model has constant velocity in all dimensions. It can be represented by

$$\mathbf{v}_n^{(\mathrm{FF})} = \mathbf{c}^{(\mathrm{FF})} \quad (6)$$

where $\mathbf{c}^{(\mathrm{FF})}$ is a vector of constant velocities. Comparing this with (3), there are several subfamilies of the Gauss–Markov model that generate movement patterns the same as those described by the fluid-flow model. They are $\{\overline{\alpha} = \mathbf{1}, \mathbf{v}_0 = \mathbf{c}^{(\mathrm{FF})}\}$, $\{\overline{\alpha} = \mathbf{0}, \overline{\mu} = \mathbf{c}^{(\mathrm{FF})}, \overline{\sigma} = \mathbf{0}\}$, and $\{\overline{\mu} = \mathbf{c}^{(\mathrm{FF})}, \overline{\sigma} = \mathbf{0}, \mathbf{v}_0 = \mathbf{c}^{(\mathrm{FF})}\}$.

iii) The random way-point model is the most prevalent mobility model in mobile ad hoc network (MANET) simulations. In this model, a mobile cycles through two phases: constant velocity and motionless. In each cycle, it randomly selects a destination, moves toward the destination at a randomly selected velocity until it reaches the destination, stays at the destination for a randomly chosen amount of time, and starts a new cycle by randomly selecting a new destination. Therefore, the movement of a mobile under this model consists of segments of fluid-flow with various level of velocity (including zero velocity).

Since the stochastic process of mobile velocity under the random way-point model is nonstationary, it cannot be obtained by *direct* application of the Gauss–Markov model. However, with the Gauss–Markov model, one only needs to employ an additional random number generator to generate movement patterns that follow the random way-point model description. This random number generator determines the duration of each segment of movement. In each segment, the Gauss–Markov model, as shown in *ii)*, generates fluid-flow-like movement patterns.

B. Predictive Distance-Based Mobility Management

In what follows, we describe the predictive mobility management scheme in the context of 1-D and 2-D systems. This can easily be extended to systems of three dimensions, such as those in an office building, for example.

---

[3]A comparative study of various mobility models, including random-walk, Gauss–Markov, and random way-point, in the context of mobile ad hoc network routing, can be found in [16].
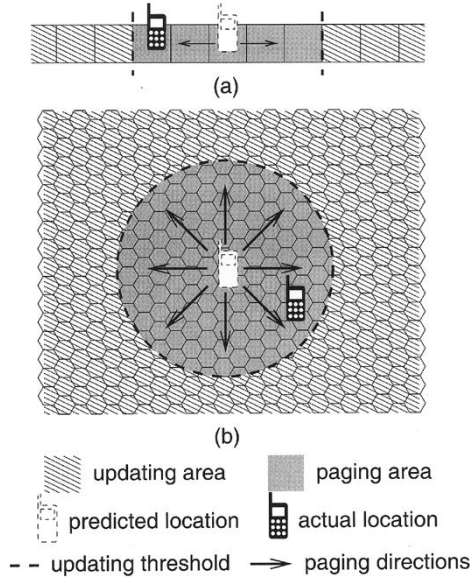
Fig. 1. Predictive location updating and selective paging. (a) 1-D system. (b) 2-D system.

1) Predictive Location Updating and Selective Paging: As explained in Section I, a mobile cannot monitor its location continuously. Here, we assume that location inspections are performed by the mobiles periodically, at every m units of time.

The proposed *predictive* distance-based scheme has the following components.

• *Prediction*: The PCS network makes predictions of the mobile's location $\mathbf{p}_n$, at time instants of mobile location inspections and call arrivals to the mobile, based on the history of the mobile's reported location and velocity. The mobile is made aware of the network's prediction, either by information transfer from the network during the last location updating or by local, concurrent computation using the same prediction algorithm as the network.

• *Updating*: At the kth location inspection instant, a mobile observes its current location $\mathbf{s}_{km}$. It transmits a location update to the PCS network if $|\mathbf{s}_{km} - \mathbf{p}_{km}|$ is greater than the distance threshold.

• *Paging*: When a call is made to a mobile at time n, the system pages the mobile in cells at and around the predicted location $\mathbf{p}_n$, in a shortest-distance-first order, until the mobile is found.[4]

Fig. 1 illustrates the predictive updating and paging schemes for the 1-D and 2-D cases, respectively.

In paging a mobile, there exists a tradeoff between the paging traffic bandwidth usage and the time delay in locating the mobile (e.g., a full-system paging scheme incurs maximum bandwidth usage and minimum delay). There is a spectrum of means to trade off between bandwidth usage and paging delay. In the rest of this paper, for both the predictive and the nonpredictive cases, we have assumed a sequential ring-by-ring paging scheme, where all cells of equal distance to the predicted location or the last reported location, respectively, are paged at the same time. For systems with different bandwidth and delay

---

[4]A time or distance threshold may be imposed, beyond which the paging terminates, in case the sought mobile has detached from the network.

---

constraint requirements, the proposed predictive scheme can be easily extended using the methods similar to those reported in [3], [14], and [24].

The updated mobile location and velocity are assumed to be contained in location databases residing in the wireline network (e.g., Home Location Registers). The reliability issues of location databases are not addressed in this paper. For reliability considerations, methods similar to those described in [12] could be employed.

2) Location Prediction: Next, we derive a formula for the predicted location of a mobile as a function of time, when its last reported location and velocity are given. Let $n = 0$ be the initial time when a mobile last updates its location and velocity. We can recursively expand (4) to express $\mathbf{v}_n$ explicitly in terms of the initial velocity $\mathbf{v}_0$

$$\mathbf{v}_n = \alpha^n \mathbf{v}_0 + (1 - \alpha^n)\mu + \sigma\sqrt{1 - \alpha^2} \sum_{i=0}^{n-1} \alpha^{n-i-1}\mathbf{w}_i. \quad (7)$$

Defining $\mathbf{s}_0 = 0$ and $\mathbf{s}_n = \mathbf{s}_{n-1} + \mathbf{v}_{n-1}$, we have

$$\mathbf{s}_n = \sum_{i=0}^{n-1} \mathbf{v}_i = \frac{1 - \alpha^n}{1 - \alpha}\mathbf{v}_0 + \left(1 - \frac{1 - \alpha^n}{1 - \alpha}\right)\mu$$

$$+ \sigma\sqrt{1 - \alpha^2} \sum_{i=0}^{n-1} \sum_{j=0}^{i-1} \alpha^{i-j-1}\mathbf{w}_j. \quad (8)$$

It is easy to show that any linear combination of $\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_n$ is Gaussian. Hence, $\mathbf{s} \triangleq \{\mathbf{s}_n\}$ is a Gaussian process.

Let $\mathbf{z}_n$ be the displacement of a mobile at time n from its last updated location, given that no location updating is performed up to time $n - 1$, i.e.,

$$|\mathbf{s}_{km} - \mathbf{p}_{km}| \leq N \quad (9)$$

for $k = 1, 2, \ldots, \lceil n/m \rceil - 1$, where N is the updating distance threshold. The joint PDF of $\mathbf{z} \triangleq \{\mathbf{z}_n\}$ is

$$f_{\mathbf{z}}(\boldsymbol{\xi}) = \begin{cases} \dfrac{f_{\mathbf{s}}(\boldsymbol{\xi})}{\int_{|\boldsymbol{\xi}_{km} - \mathbf{p}_{km}| \leq N} f_{\mathbf{z}}(\boldsymbol{\xi})\, d\boldsymbol{\xi}}, \\ \qquad\qquad |\boldsymbol{\xi}_{km} - \mathbf{p}_{km}| \leq N, \\ \qquad\qquad k = 1, 2, \ldots, \lceil n/m \rceil - 1; \\ 0, \qquad \text{otherwise} \end{cases} \quad (10)$$

where $\boldsymbol{\xi} \triangleq \{\boldsymbol{\xi}_n\}$, and $f_{\mathbf{s}}(\boldsymbol{\xi})$ is the joint PDF of $\mathbf{s}$.

Since $\{\mathbf{s}_n\}$ is a Gaussian process, the maximum-likelihood estimation of $\mathbf{s}_n$ is $\mathrm{E}[\mathbf{s}_n]$. Due to symmetry in the definition of $\{\mathbf{z}_n\}$, as shown in (10), the maximum-likelihood estimation of $\mathbf{z}_n$ is

$$\mathbf{p}_n = \mathrm{E}[\mathbf{z}_n] = \mathrm{E}[\mathbf{s}_n]$$

$$= \frac{1 - \alpha^n}{1 - \alpha}\mathbf{v}_0 + \left(1 - \frac{1 - \alpha^n}{1 - \alpha}\right)\mu. \quad (11)$$

Furthermore, since the PDF of $\mathbf{z}_n$ is concave and symmetrical, the proposed shortest-distance-first paging scheme conforms with the highest-probability-first rule for optimal selective paging. Therefore, we can expect the predictive scheme to

incur lower mobility management cost than the nonpredictive distance-based scheme.

C. Gauss–Markov Model Parameter Estimation

In a practical system, the mobility pattern can be approximated by the Gauss–Markov model with varying mean, variance, and memory level. Applying the algorithm in Section II-B, the movement of a mobile is predicted based on the estimated parameters $\hat{\mu}$, $\hat{\sigma}$, and $\hat{\alpha}$, obtained at the previous location updates. Therefore, the accurate location prediction depends upon the accurate dynamic estimation of the Gauss–Markov parameters, as well as the degree of variations in these parameters between location updates.

The topic of dynamic parameter estimation has seen many contributions within the signal-processing community. In this work, we propose a simple estimator that computes the mean, variance, and memory level of a mobile, *separately*, in each dimension. We define a parameter estimation window w_{est}. In practice, w_{est} is chosen such that it is small enough for the mobile movement to remain approximately stationary but large enough to provide sufficient amount of data samples. Given the previous w_{est} samples of the mobile velocity $\{v_i\}$, the Gauss–Markov parameters are estimated by

$$\hat{\mu} = \frac{1}{w_{est}} \sum_{i=1}^{w_{est}} v_i \qquad (12)$$

$$\hat{\sigma}^2 = \frac{1}{w_{est}-1} \sum_{i=1}^{w_{est}} (v_i - \hat{\mu})^2 \qquad (13)$$

$$\hat{\sigma}_1^2 = \frac{1}{w_{est}-1} \sum_{i=1}^{w_{est}-1} (v_i - \hat{\mu})(v_{i+1} - \hat{\mu}) \qquad (14)$$

$$\hat{\alpha} = \begin{cases} 1, & \text{if } \hat{\sigma} \approx 0 \\ \max\left\{0, \frac{\hat{\sigma}_1^2}{\hat{\sigma}^2}\right\}, & \text{otherwise.} \end{cases} \qquad (15)$$

In the above estimation, both $\hat{\sigma}^2$ and $\hat{\sigma}_1^2$ are biased because of the correlation between elements in $\{v_i\}$. Despite this drawback, there are two reasons that we have chosen this algorithm. First, this estimator is simple to implement in practice. Second, a suboptimal parameter estimator does *not* invalidate the correctness of the proposed predictive location updating and paging scheme.

With correlated data and limited estimation window size, an optimal estimator has not been found. However, it is stated in [17] that the above estimator can outperform an unbiased estimator.

Thus, the Gauss–Markov parameter estimates of each mobile are recomputed at each of its location updating events. These estimates are used by the mobile and the system to predict the future location of the mobile. Section IV-C demonstrates the effectiveness of combining the above estimator with distance-based predictive mobility management.

III. COST EVALUATION OF THE PREDICTIVE DISTANCE-BASED MOBILITY MANAGEMENT SCHEME

In this section, we describe an analytical framework to evaluate the mobility management cost of the predictive distance-based scheme, assuming ideal Gauss–Markov mobility pattern. The performance of the proposed scheme in the nonideal time-varying case is studied in Section IV.

The total cost of mobility management *per call arrival* is defined as

$$C_{total} = C_{inspect} + C_{update} + C_{page} \qquad (16)$$

where $C_{inspect}$, C_{update}, and C_{page} are the average location inspection cost per call arrival, the average updating cost per call arrival, and the average paging cost per call arrival, respectively.

Clearly, there exists a tradeoff between location updating and mobile paging. The more often a mobile terminal updates its location, the more precise is the network's knowledge of its location when a call arrives, which could lead to lower paging cost. However, location updating drains the limited energy supply of a mobile terminal, takes away part of the precious wireless channel capacity, and induces access to potentially far away and heavily loaded location databases. Therefore, the rate of location updating, which depends on the location updating threshold in this case, should be optimized to reduce the total cost.

The frequency of location inspection is another design parameter that needs to be optimized. On the one hand, location inspection takes up the limited processor power and energy supply of a mobile terminal. On the other hand, it improves the mobile's awareness of its location for timely location updating, which is particularly important when the frequency of call arrivals is high.

In order to determine the optimal location inspection frequency and the optimal location updating threshold, we propose the following framework for evaluating the cost of mobility management. We first describe the cost evaluation framework of a 1-D system. We then extend that to 2-D and 3-D systems, and follow that with an approximation scheme, which reduces the computational complexity of multidimensional cost evaluation to that of a 1-D system.

A. 1-D Cost Evaluation

Since the position of a mobile (e.g., the residence cell identity) is closely monitored by the network within a phone call duration, a call arrival has the same effect as a mobile location update. To distinguish the two types of location updating, we refer to a location update based on distance as *autonomous update*. Furthermore, we can view a call duration as a point in time, where a location update is performed, and a new cycle of location inspection, location updating, and mobile paging starts.

Suppose calls arrive with the interarrival time PDF $f_{call}(n)$. The average location inspection cost per call is

$$C_{inspect} = C_i \sum_{n=0}^{\infty} f_{call}(n) \lfloor n/m \rfloor \qquad (17)$$

where C_i denotes the cost of a single location inspection. If we assume that the call arrival times are exponentially distributed with arrival rate λ, we can simplify (17) to

$$C_{inspect} = C_i \frac{(1-\lambda)^m}{1-(1-\lambda)^m}. \qquad (18)$$

The detailed derivation is presented in Appendix A.

To determine C_{update}, we first consider the time interval between two consecutive autonomous location updates without the interruption of phone calls. We compute the PDF of time between a location update (autonomous update or call-arrival update) and the next autonomous update, denoted $f_{\text{upd}}(n)$.

Let η_n and z_n be the velocity and displacement of a mobile at time n, respectively, given that no location updating is performed up to the location inspection at time $km = \lfloor n/m \rfloor m$. Shifting the center of the PDFs of η_n and z_n to the origin, we define

$$u_n = \eta_n - \mu_{\eta_n} \tag{19}$$

$$r_n = z_n - \mu_{z_n}. \tag{20}$$

Then, we have $u_0 = r_0 = 0$, and for $n \neq km$

$$u_n = \alpha u_{n-1} + \sigma\sqrt{1 - \alpha^2}\, w_{n-1} \tag{21}$$

and

$$r_n = r_{n-1} + u_{n-1}. \tag{22}$$

We further define u'_{km} and r'_{km} as the mean-shifted velocity and displacement of a mobile at the kth location inspection, respectively, given that no location updating is performed up to the $(k-1)$th location inspection (i.e., r'_{km} could be greater than N). Obviously, u_{km} is u'_{km} conditioned on $|r'_{km}| \leq N$, and r_{km} is r'_{km} conditioned on $|r'_{km}| \leq N$.

In order to compute the probability of an update at the kth location inspection, we need to compute the PDF of r_{km}. It's easy to see that the statistics of (r_{km}, u_{km}) is completely determined by $(r_{(k-1)m}, u_{(k-1)m})$. Therefore, we can compute the joint PDF $f_{r_{km}, u_{km}}(r, u)$, through a recursion on k as follows.

Let

$$y_k = \sum_{i=km}^{(k+1)m-1} u_i. \tag{23}$$

Then

$$r'_{(k+1)m} = r_{km} + y_k. \tag{24}$$

By the Markovian property of u_n, we have

$$
\begin{aligned}
&f_{y_k|(u_{km}, u'_{(k+1)m})}(y, \xi_1, \xi_2) \\
&= f_{y_0|(u_0, u'_m)}(y, \xi_1, \xi_2) \\
&= \frac{1}{\sigma_{y_0|u_0}\sqrt{2\pi(1-\gamma^2)}} \\
&\quad \cdot \exp\left\{ -\frac{\left[(y - \mu_{y_0|u_0}) - \gamma\frac{\sigma_{y_0|u_0}}{\sigma_{u'_m|u_0}}(\xi_2 - \mu_{u'_m|u_0}) \right]^2}{2\sigma_{y_0|u_0}^2(1-\gamma^2)} \right\}
\end{aligned}
\tag{25}
$$

where

$$\mu_{u'_m|u_0} = \alpha^m \xi_1 \tag{26}$$

$$\mu_{y_0|u_0} = \frac{1-\alpha^m}{1-\alpha}\xi_1 \tag{27}$$

$$\sigma_{u'_m|u_0}^2 = (1 - \alpha^{2m})\sigma^2 \tag{28}$$

$$\sigma_{y_0|u_0}^2 = \frac{-(1-\alpha^m)^2 + m(1-\alpha^2) - 2\alpha(1-\alpha^m)}{(1-\alpha)^2}\sigma^2 \tag{29}$$

$$\gamma = \frac{C_{(u'_m, y_0)|u_0}}{\sigma_{u'_m|u_0}\,\sigma_{y_0|u_0}} \tag{30}$$

and

$$C_{(u'_m, y_0)|u_0} = \frac{(1-\alpha^m)(\alpha - \alpha^m)}{1-\alpha}\sigma^2. \tag{31}$$

The derivation of $f_{y_0|(u_0, u'_m)}(y, \xi_1, \xi_2)$ is presented in Appendix B.

Since r_{km} and y_k are independent given u_{km}, we have from (24)

$$
\begin{aligned}
&f_{r'_{(k+1)m}|(u_{km}, u'_{(k+1)m})}(r, \xi_1, \xi_2) \\
&= f_{r_{km}|u_{km}}(r, \xi_1) * f_{y_{km}|(u_{km}, u'_{(k+1)m})}(r, \xi_1, \xi_2) \\
&= f_{r_{km}|u_{km}}(r, \xi_1) * f_{y_0|(u_0, u'_m)}(r, \xi_1, \xi_2) \tag{32}
\end{aligned}
$$

where $*$ denotes linear 1-D convolution over the variable r.

Then, by Bayes' rule and taking the marginal density, we have

$$
\begin{aligned}
&f_{r'_{(k+1)m}, u'_{(k+1)m}}(r, \xi_2) \\
&= \int_{-\infty}^{\infty} f_{r'_{(k+1)m}|(u_{km}, u'_{(k+1)m})}(r, \xi_1, \xi_2) \\
&\qquad \cdot f_{u_{km}, u'_{(k+1)m}}(\xi_1, \xi_2)\, d\xi_1 \\
&= \int_{-\infty}^{\infty} f_{r'_{(k+1)m}|(u_{km}, u'_{(k+1)m})}(r, \xi_1, \xi_2) \\
&\qquad \cdot f_{u'_{(k+1)m}|u_{km}}(\xi_2, \xi_1) \\
&\qquad \cdot \int_{-\infty}^{\infty} f_{r_{km}, u_{km}}(\zeta, \xi_2)\, d\zeta\, d\xi_1 \tag{33}
\end{aligned}
$$

where

$$
\begin{aligned}
&f_{u'_{(k+1)m}|u_{km}}(\xi_2, \xi_1) \\
&= f_{u'_m|u_0}(\xi_2, \xi_1) \\
&= \frac{1}{\sigma_{u'_m|u_0}\sqrt{2\pi}}\exp\left[-\frac{(\xi_2 - \mu_{u'_m|u_0})^2}{2\sigma_{u'_m|u_0}^2} \right]. \tag{34}
\end{aligned}
$$

Because of the distance-based updating, we obtain

$$f_{r_{(k+1)m}, u_{(k+1)m}}(r, u) = \frac{f_{r'_{(k+1)m}, u'_{(k+1)m}}(r, u)P_N(r)}{\int_{-\infty}^{\infty}\int_{-N}^{N} f_{r'_{(k+1)m}, u'_{(k+1)m}}(r, u)\, dr\, du} \tag{35}$$

where

$$P_N(r) = \begin{cases} 1, & -N \leq r \leq N \\ 0, & \text{otherwise.} \end{cases} \tag{36}$$

Let $R_{\text{upd}}(k)$ be the probability that there is no update up to and including the kth location inspection instant. We have $R_{\text{upd}}(0) = 1$, and

$$R_{\text{upd}}(k+1)$$
$$= R_{\text{upd}}(k) \int_{-\infty}^{\infty} \int_{-N}^{N} f_{r'_{(k+1)m} u'_{(k+1)m}}(r, u) \, dr \, du. \quad (37)$$

The above iteration to compute $f_{r_{km}}(r)$ can be terminated either when $R_{\text{upd}}(k)$ is sufficiently small, or when the series of PDF's $f_{r_{km}, u_{km}}(r, u)$ converges, e.g., in terms of normalized squared error

$$\frac{\int_{-\infty}^{\infty} \int_{-N}^{N} [f_{r_{km}, u_{km}}(r, u) - f_{r_{(k-1)m}, u_{(k-1)m}}(r, u)]^2 \, dr \, dv}{\int_{-\infty}^{\infty} \int_{-N}^{N} [f_{r_{km}, u_{km}}(r, u)]^2 \, dr \, dv} < \epsilon$$

where ϵ is a predefined threshold. Suppose the iteration is terminated at the qth step. Since $f_{r_{km}, u_{km}}(r, u)$ uniquely determines $f_{r'_{(k+1)m}, u'_{(k+1)m}}(r, u)$, for $k \geq q$, $R_{\text{upd}}(k+1)/R_{\text{upd}}(k)$ is essentially constant and

$$R_{\text{upd}}(k+1) \approx \frac{R_{\text{upd}}(q)}{R_{\text{upd}}(q-1)} R_{\text{upd}}(k). \quad (38)$$

Thus, we can compute $R_{\text{upd}}(k)$ successively for all $k \geq q$. We carry out this computation until $R_{\text{upd}}(k)$ is small enough.

Thus, we obtain the PDF of the time between a location update to the next autonomous update by

$$f_{\text{upd}}(km) = R_{\text{upd}}(k-1) - R_{\text{upd}}(k), \qquad k = 1, 2, \ldots \quad (39)$$

and $f_{\text{upd}}(n) = 0$ for n not a integer multiple of m. Since $f_{\text{upd}}(km)$ is not dependent on the history of a mobile's location or velocity before time $k = 0$, the time between consecutive autonomous updates is independent and identically distributed (i.i.d.).

Next, we consider the location updates between two successive call arrivals. Since a call arrival has the same effect as a location update, the time between the first call arrival and the first autonomous update has PDF $f_{\text{upd}}(n)$. The i.i.d. location update time intervals comprise a renewal process with the PDF $f_{\text{upd}}(n)$.

Let $U(n)$ denote the number of location updates within the time interval of length n between two successive call arrivals. Then

$$\Pr[U(n) = i] = F^{(i)}(n) - F^{(i+1)}(n) \quad (40)$$

where

$$F^{(i)}(n) \triangleq \sum_{l=0}^{n} f^{(i)}(l) \quad (41)$$

and $f^{(i)}(n)$ is the PDF of time between a location update and the ith autonomous update in the future, as given by the following recursive relation:

$$f^{(1)}(n) = f_{\text{upd}}(n) \quad (42)$$
$$f^{(i)}(n) = f^{(i-1)}(n) * f_{\text{upd}}(n) \quad (43)$$

where $*$ denotes the discrete convolution.

Let $M(n)$ be the expected value of $U(n)$. Then

$$M(n) = \sum_{i=1}^{\infty} i(F^{(i)}(n) - F^{(i+1)}(n))$$
$$= \sum_{i=1}^{\infty} i F^{(i)}(n) - \sum_{i=2}^{\infty} (i-1) F^{(i)}(n)$$
$$= \sum_{i=1}^{\infty} F^{(i)}(n). \quad (44)$$

Using (44), the average updating cost per call arrival can be obtained by

$$C_{\text{update}} = C_u \sum_{n=0}^{\infty} f_{\text{call}}(n) M(n) \quad (45)$$

where C_u is the cost of a single autonomous location update.

To find C_{page}, we first need to determine the distribution of r_n, for $n = km + 1, km + 2, \ldots, km + m - 1$, given $f_{r_{km}, u_{km}}(r, u)$. Let

$$\delta_{k, i} = \sum_{j=km}^{km+i-1} u_{km+j}, \qquad i = 1, 2, \ldots, m. \quad (46)$$

Similar to the mean and variance of $y_0|v_0$ derived in Appendix B, we can show that $\delta_{k, i}|u_{km}$ is Gaussian with mean $((1 - \alpha^i)/(1 - \alpha))u_{km}$ and variance $((-(1 - \alpha^i)^2 + i(1 - \alpha^2) - 2\alpha(1 - \alpha^i))/((1 - \alpha)^2))\sigma^2$. Then, since

$$r_{km+i} = r_{km} + \delta_{k, i} \quad (47)$$

and r_{km} and $\delta_{k, i}$ are mutually independent given u_{km}, we have

$$f_{r_{km+i}}(r) = \int_{-\infty}^{\infty} f_{r_{km}, u_{km}}(r, u) * f_{\delta_{k, i}|u_{km}}(r, u) \, du \quad (48)$$

where $*$ denotes convolution in terms of r.

Let $K_p(r)$ be the number of cells in which the network needs to page a mobile when the mobile is at a displacement of r from its predicted location. In a 1-D predictive system with a shortest-distance-first paging rule

$$K_p(r) = 2\lfloor r/S + 0.5 \rfloor + 1 \quad (49)$$

where S is the cell size. Then, the expected number of cells paged when a call arrives at time i from the last location updating is

$$g(i) = \int_{-\infty}^{\infty} K_p(r) f_{r_i}(r) \, dr. \quad (50)$$

Let $P(n)$ be the expected number of cells paged when a call arrives at time n from the last call arrival. We can compute $P(n)$ with the following recursive equation:

$$P(n) = R_{\text{upd}}(\lfloor n/m \rfloor) g(n) + f_{\text{upd}}(m) P(n - m)$$
$$+ f_{\text{upd}}(2m) P(n - 2m) + \cdots$$
$$+ f_{\text{upd}}(\lfloor n/m \rfloor m) P(n - \lfloor n/m \rfloor m). \quad (51)$$

Then, the average paging cost per call arrival is given by

$$C_{\text{page}} = C_p \sum_{n=0}^{\infty} f_{\text{call}}(n) P(n) \tag{52}$$

where C_p is the cost of paging a single cell.

B. 2-D Cost Evaluation

Section III-A describes the analytical framework for evaluating the predictive mobility management scheme in a 1-D system. In a 2-D system, similarly to (19) and (20), we can define mean-shifted velocity $\mathbf{u}_n = (u_n^x, u_n^y)^T$ and mean-shifted displacement $\mathbf{r}_n = (r_n^x, r_n^y)^T$, given that no location updating is performed up to the $\lfloor n/m \rfloor$th location inspection. Then, similarly to (19)–(39), we can compute $f_{\text{upd}}^{(2d)}(n)$ and $f_{r_{km}^x, r_{km}^y, u_{km}^x, u_{km}^y}(r^x, r^y, u^x, u^y)$ through an iterative procedure over $f_{\mathbf{r}_{km}, \mathbf{u}_{(k-1)m}, \mathbf{u}_{km}}(\mathbf{r}, \mathbf{u}_1, \mathbf{u}_2)$. Here, the location cut-off function, as shown in (36), becomes

$$P_N^{(2d)}(r^x, r^y) = \begin{cases} 1, & \sqrt{(r^x)^2 + (r^y)^2} \leq N \\ 0, & \text{otherwise} \end{cases} \tag{53}$$

where N is the threshold radius for location updating.

Then, the computation of C_{inspect} remains unchanged and the computation of C_{update} can be carried out exactly as in (40)–(45). The computation of C_{page} is similar to that described in (46)–(52), except that $K_p(\mathbf{r})$ in (49) needs to be redefined.

The number of cells paged by the network to find a mobile is largely dependent upon the cell shape and lattice arrangement. In order to obtain analytical results that are compatible with all cell configurations, we assume that the number of cells paged is proportional to the size of the area paged by the network. Since the optimal, shortest-distance-first and ring-by-ring paging pattern is used, the area paged when the mobile is at (r^x, r^y) is a circle centered at the predicted location and with radius $\sqrt{(r^x)^2 + (r^y)^2}$. Thus, here the 2-D $K_p^{(2d)}(\mathbf{r})$ is defined as

$$K_p^{(2d)}(r^x, r^y) = \left\lfloor \pi \left((r^x)^2 + (r^y)^2 \right) / S^{(2d)} \right\rfloor + 1 \tag{54}$$

where $S^{(2d)}$ is the average area of a cell.

As hinted by (32)–(34), the direct extension of the 1-D algorithm to the 2-D case can potentially be very computationally intensive, since it requires convolution and other manipulation of $f_{\mathbf{r}_{km}, \mathbf{u}_{(k-1)m}, \mathbf{u}_{km}}(\mathbf{r}, \mathbf{u}_1, \mathbf{u}_2)$ with six free variables. In what follows, we introduce an approximating algorithm for 2-D cost evaluation, which requires only as much computation as the 1-D algorithm.

C. 2-D Cost Approximation

As shown in Section III-B, the difficulty there is in computing the joint PDF of location and velocity in both dimensions. This problem can be alleviated by an approximation based on the assumption that a mobile performs location updating separately in each of the two dimensions. In this case, we define a new location cut-off function

$$P_L^{(2d)}(r^x, r^y) = \begin{cases} 1, & -L \leq r^x \leq L \text{ and } -L \leq r^y \leq L \\ 0, & \text{otherwise} \end{cases} \tag{55}$$

where L is the threshold distance for location updating in each dimension, and

$$L = \sqrt{\pi N^2} \tag{56}$$

which maintains the same size of the range of possible mobile location immediately after a location inspection.

With this location updating assumption, the mobile location and update time statistics can be computed independently in each dimension, and then combined to give the 2-D statistics. The 2-D autonomous update time is the minimum of the 1-D autonomous update time in each dimension. In particular, since we have assumed that the mobiles have the same and independent mobility pattern in both dimensions, we have

$$R_{\text{upd}}^{(2d)}(k) = R_{\text{upd}}^2(k) \tag{57}$$

and hence

$$f_{\text{upd}}^{(2d)}(km) = 2 f_{\text{upd}}(km) R_{\text{upd}}(km) + f_{\text{upd}}^2(km). \tag{58}$$

The 2-D mobile location PDF can be obtained by

$$f_{r_n^x, r_n^y}(r^x, r^y) = f_{r_n}(r^x) f_{r_n}(r^y). \tag{59}$$

Then, we can use the same procedures described in Section III-B to compute the updating and paging costs. To make the approximation results more precise, we can modify $K_p^{(2d)}(r^x, r^y)$ to reflect the mobile location constraint introduced by (55). For simplicity, in this work, we only consider this constraint at time $n = km$. In this case, the paging area is the intersection of a circle and a square centered at the predicted location, with radius $N = \sqrt{(r^x)^2 + (r^y)^2}$ and side-length $2L$, respectively. Thus, we redefine

$$
K_p^{(2d)}(r^x, r^y)
$$
$$
= \begin{cases}
\left\lfloor \dfrac{\pi N^2}{S^{(2d)}} \right\rfloor + 1, & N \leq L \\[2ex]
\left\lfloor \dfrac{\pi N^2}{S^{(2d)}} - \dfrac{4}{S^{(2d)}} \left(N^2 \tan^{-1} \dfrac{\sqrt{N^2 - L^2}}{L} \right. \right. \\[2ex]
\qquad \left. \left. - L\sqrt{N^2 - L^2} \right) \right\rfloor + 1, & N > L.
\end{cases} \tag{60}
$$

Since (55) represents a suboptimal updating scheme, C_{total} obtained here is an upper bound of the actual cost that can be obtained by the direct 2-D computation in Section III-B. In all experiments that we have conducted, this upper bound is within 2% of the actual cost. Thus, we conclude that the approximation is quite reasonable, allowing us to significantly reduce the computation complexity.

IV. NUMERICAL RESULTS AND COMPARISONS

In this section, we evaluate the cost of a 2-D system. When analyzing the cost of the predictive scheme in the case of ideal Gauss–Markov mobility pattern, we use the upper-bound approximation described in Section III-C to expedite the computations. In the case of time-varying mobility pattern, the performance of the predictive scheme, in conjunction with the proposed parameter estimator, is demonstrated through simulation.

In order to reduce the redundancy of numerical representation, we scale both the spatial parameters and the cost parame-

ters. We normalize all distance and velocity values according to the cell size, such that the area of a cell is assumed to be one. Furthermore, since only the *relative* costs of location inspection, location updating, and mobile paging affect the performance optimization, we also normalize all costs to have the units of C_p, the cost of a single paging. Thus, C_u now represents the ratio of the cost of one location update to the cost of a single paging, and C_i now represents the ratio of the cost of one location inspection to the cost of a single paging.

The costs C_i, C_u, and C_p depend on multiple factors such as the mobile battery and processor capacity, the system wireless bandwidth, and the location database architecture. It is impossible to quantify all these factors using a single metric. In fact, it is difficult to properly quantify any one of these three costs with a single metric. Therefore, the values of C_i and C_u in our analysis merely represent the relative importance of location inspection, location updating, and mobile paging in evaluating the performance of a system.

Clearly, the relative importance of these different costs is determined by the system design details on a case-by-case basis. Furthermore, these costs may change over time, depending on the time varying characteristics of the battery power, wireless channel usage, database loading, network congestion, and many other factors that a system designer need to consider. In this paper, we consider the broad applicability of predictive mobility management in different networks. Therefore, instead of giving the numerical results of one specific type of network, we study the optimization and performance of the proposed scheme with a wide range of C_i and C_u values.

For the case of ideal Gauss–Markov mobility pattern, we are interested in understanding how C_u, C_i, and the other system variables, namely, the memory factor exponent β, the average velocity μ, the standard deviation σ, the location inspection period m, and the updating distance N, affect the performance of the predictive scheme. In particular, we are interested in the joint optimization of m and N, namely

$$(m_{\text{opt}}, N_{\text{opt}}) = \arg \min_{m, N} C_{\text{total}}(m, N, \beta, \mu, \sigma, \lambda, C_u, C_i) \tag{61}$$

which gives the minimum cost of mobility management $C_{\text{total}}(m_{\text{opt}}, N_{\text{opt}})$.

We are also interested in the performance gain that the predictive scheme can achieve over the nonpredictive distance-based scheme. In particular, we are interested in how the cost ratio is affected by the aforementioned system parameters.

Although our analytical framework accommodates a general call-arrival distribution, in the following numerical computations, we assume that the interarrival time of calls is exponentially distributed with arrival rate λ [calls/unit time].

A. Joint Optimization of m and N With Ideal Gauss–Markov Mobility Pattern

Assuming ideal prediction, μ has no effect on the cost of predictive mobility management. Other than that, however, it is not easy to solve (61) using general numerical optimization tools. Therefore, in what follows, given a set of system parameters, we first compute C_{total} for a range of values of m and N and then pick the pair that gives the minimum cost.
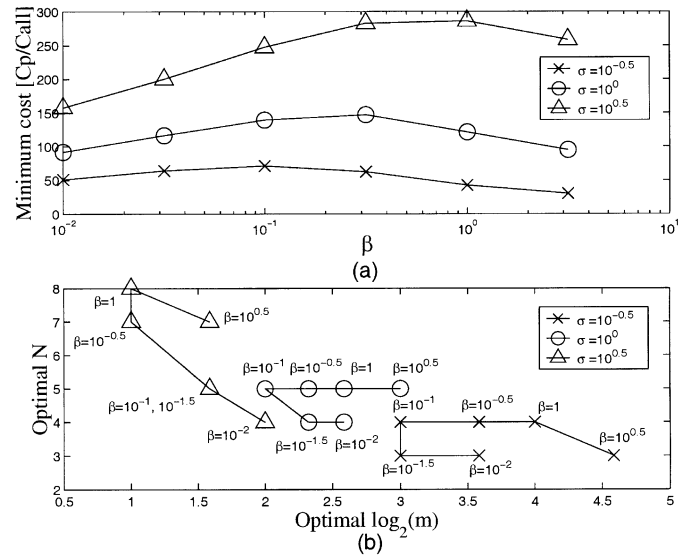

Fig. 2. Joint optimization of m and N versus β and σ, for $\lambda = 0.01$, $C_u = 10$, and $C_i = 1$. (a) Minimum cost. (b) Optimal m and N.

In Figs. 2–4, we study the effect of β, σ, λ, C_i, and C_u on the minimization of C_{total}.

Fig. 2 presents the plots of the minimum C_{total} and the corresponding $(m_{\text{opt}}, N_{\text{opt}})$, as β increases along the values $\{10^{-2}, 10^{-1.5}, 10^{-1}, 10^{-0.5}, 10^0, 10^{0.5}\}$ (i.e., α decreases along the values $\{0.99, 0.97, 0.90, 0.73, 0.37, 0.04\}$), for various values of σ and for $\lambda = 0.01$, $C_u = 10$, and $C_i = 1$.

Fig. 2(a) illustrates a seemingly counterintuitive phenomenon. It shows that the minimal total cost is not always decreasing with the memory level ($\alpha = e^{-\beta}$) of a mobile's velocity. On the one hand, for fixed σ and small β, as β increases, a mobile's velocity becomes less correlated, so the total cost of using prediction increases. On the other hand, when β is large, the variance of $\sum_{i=0}^{n} v_i$ approaches a local minimum $n\sigma^2$, so a mobile's position becomes more predictable, and therefore the total cost decreases.

In addition, Fig. 2(a) shows that the minimal total cost is an increasing function of the variance of a mobile's velocity. This matches the intuition that the cost of predictive mobility management increases as a mobile's velocity becomes more chaotic.

Fig. 2(b) shows that, corresponding to the results in Fig. 2(b), the optimal m and the optimal N are, respectively, convex and concave functions of β. However, the optimal m is always a decreasing function of σ, and the optimal N is always an increasing function of σ. Therefore, as a mobile's movement becomes more unpredictable, more frequent location inspections and longer update threshold distance should be used.

Fig. 3 presents the plots of the minimum C_{total} and the corresponding $(m_{\text{opt}}, N_{\text{opt}})$, for various values of λ and C_i. Here, C_u is fixed at 10. Fig. 4 presents the plots of the minimum C_{total} and the corresponding $(m_{\text{opt}}, N_{\text{opt}})$, for various values of λ and C_u. Here, C_i is fixed at 1. In both of these figures, $\beta = 10^{-0.5}$ (i.e., $\alpha = 0.73$), and $\sigma = 0.5$.

The plots in Figs. 3(a) and 4(a) demonstrate that the minimum cost of mobility management per call arrival is approximately an exponentially decreasing function of the call arrival probability. Since a mobile's location is tracked during a call, as the interar-
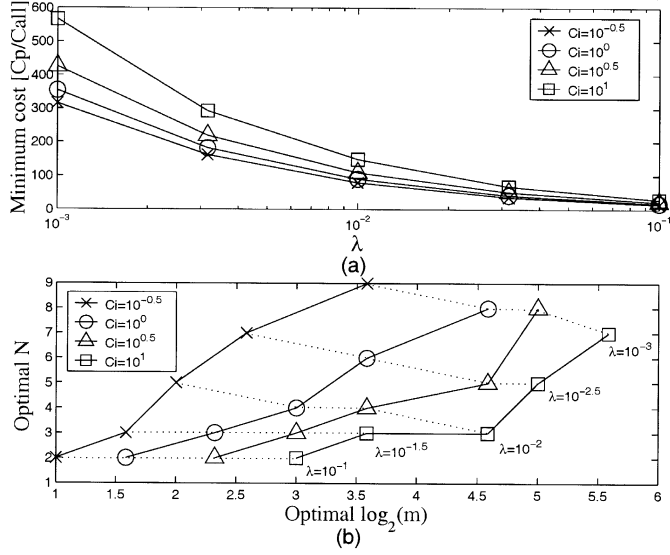
Fig. 3. Joint optimization of m and N versus λ and C_i, for $\beta = 10^{-0.5}$, $\sigma = 0.5$, and $C_u = 10$. (a) Minimum cost. (b) Optimal m and N.

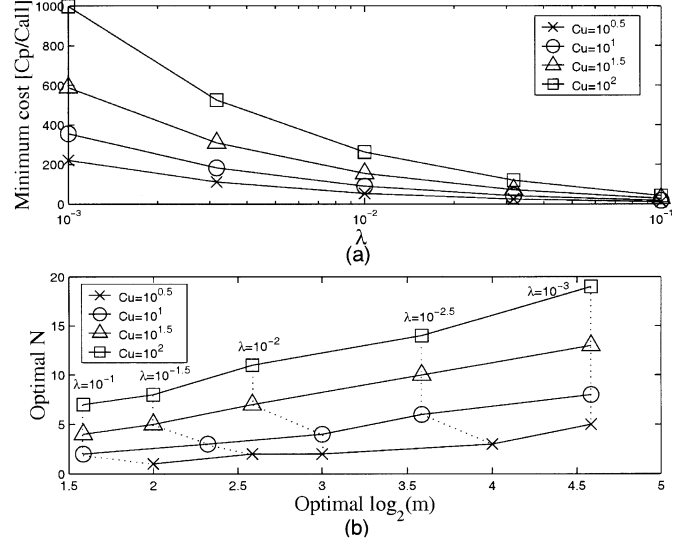

Fig. 4. Joint optimization of m and N versus λ and C_u, for $\beta = 10^{-0.5}$, $\sigma = 0.5$, and $C_i = 1$. (a) Minimum cost. (b) Optimal m and N.

rival time decreases, there is less unpredictable mobile random movement between calls, which leads to lower mobility management cost per call. As λ approaches 1 [arrivals/time unit], the total cost should approach $1[C_p]$. In this case no location inspection or location updating is necessary, and the only cost is incurred when a call arrives and the system pages once to verify that the mobile is there.

Figs. 3(b) and 4(b) demonstrate that both the optimal location inspection period and the optimal updating distance decrease as the call arrival probability increases. This is intuitive since, as call arrives more often, a mobile is better off to check and update its location more often to reduce the cost of the frequent paging operation. From these figures, we also see that m_{opt} is more sensitive to C_i than N_{opt} is, and on the other hand, N_{opt} is more sensitive to C_u than m_{opt} is. For example, when $\lambda = 0.01$ [arrivals/time unit], as shown in Figs. 3(b), as C_i goes from 1 to $10^{0.5}$, m_{opt} increases from 8 to 12, while N_{opt} remains unchanged at 4. However as shown in Fig. 4(b), as C_u goes from $10^{1.5}$ to 10^2, N_{opt} increases from 7 to 11, while m_{opt} remains unchanged at 4. The relative insensitiveness of the optimal m and N to some of the system parameters is a welcome property, especially when the proposed scheme is applied to a system where the cost factors and are time varying and need to be estimated.

B. Comparison With the Nonpredictive Distance-Based Scheme

For the nonpredictive distance-based scheme, we use computer simulations to determine its cost when the system parameters take various combinations of values. In these simulations, we assume an infinite 2-D space that is divided into cells of size 1, where a mobile travels according to the 2-D isotropic Gauss–Markov process defined by the mobility parameters β, μ, and σ. The simulations are time driven. At the time of initiation, the mobile is assumed to have just experienced a call arrival. Thus, it starts from the origin ($\mathbf{s}_0 = \mathbf{0}$), and has initial velocity with the Gaussian distribution defined by μ and σ.

The time of the next call arrival is randomly generated following the exponential distribution with arrival rate λ. Until a call arrives, the mobile inspects its position every m clock ticks. If the mobile is N or more unit of distance away from the origin, a location update is performed, and the origin is shifted to the current location. When a call arrives, the paging cost is computed based on the mobile's distance from the origin, using (54). The updating cost is computed based on counting the total number of location updates performed since time initiation. The above experiment is repeated 10^5 times, and the average is taken.

To find the joint optimization of m and N through simulations in the nonpredictive case would be too daunting a task. Instead, to facilitate the comparison between the two schemes, for both schemes, we treat m as a nondesign parameter, and only optimize N. Thus

$$N_{\text{opt}}^p = \arg\min_N C_{\text{total}}^p(N, \beta, \mu, \sigma, \lambda, C_u, C_i, m) \qquad (62)$$

$$N_{\text{opt}}^{np} = \arg\min_N C_{\text{total}}^{np}(N, \beta, \mu, \sigma, \lambda, C_u, C_i, m). \qquad (63)$$

Furthermore, since both the predictive and nonpredictive schemes have the same C_{inspect}, independent of N, we drop this term and define the performance gain

$$G(\beta, \mu, \sigma, \lambda, C_u, m) = \frac{C_{\text{update}}^{np}(N_{\text{opt}}^{np}) + C_{\text{page}}^{np}(N_{\text{opt}}^{np})}{C_{\text{update}}^p(N_{\text{opt}}^p) + C_{\text{page}}^p(N_{\text{opt}}^p)}. \qquad (64)$$

Thus, the predictive performance gain is a function of six independent variables. Instead of attempting to plot the performance gain in the six-dimensional space, we divide the variables into two groups, the mobility-related parameters (β, μ, σ) and the traffic and mobility-management parameters (λ, C_u, m). For each group of variables, we study the effect of these variable in detail, while the variables in the other group are fixed.

In Fig. 5, we study the effect of the mobility pattern, namely, (β, μ, σ), on the performance gain. The other parameters are set to (λ, C_u, m) = (0.01, 10, 10).

Fig. 5(a) presents the plots of the performance gain versus β, for various values of μ, and with fixed $\sigma = 0.5$. Here β

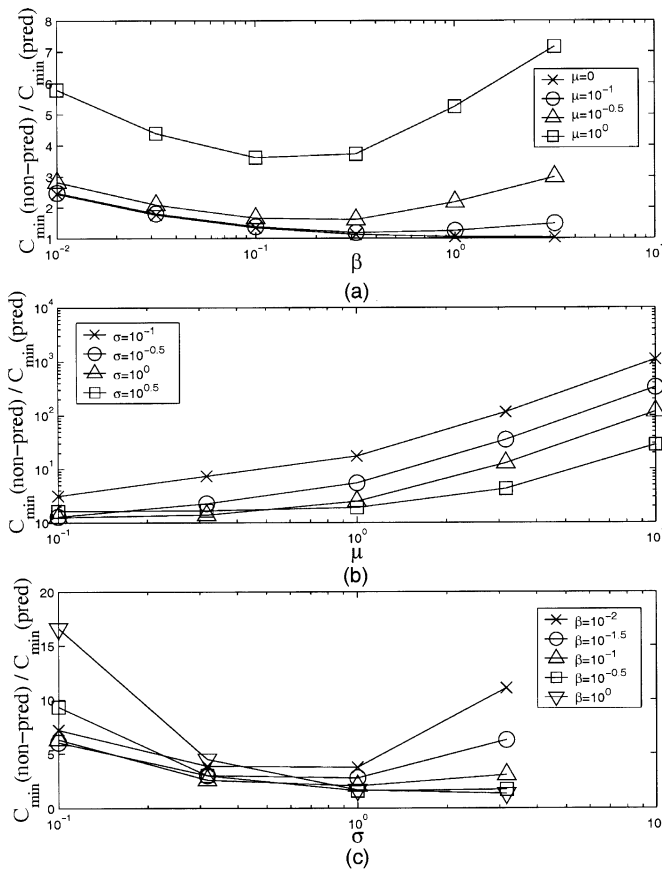Fig. 5. Performance gain with ideal Gauss–Markov mobility model, for $\lambda = 0.01$, $C_u = 10$, and $m = 10$. (a) Versus β and μ, with $\sigma = 0.5$. (b) Versus μ and σ, with $\beta = 10^{-0.05}$ ($\alpha = 0.73$). (c) Versus σ and $beta$, with $\mu = 0.5$.

Fig. 6. Performance gain with ideal Gauss–Markov mobility model, for $\mu = 0.5$, $\sigma = 0.5$, and $\beta = 10^{-0.05}$. (a) Versus λ and C_u, with $m = 10$. (b) Versus C_u and m, with $\lambda = 0.01$. (c) Versus m and λ, with $C_u = 10^{1.5}$.

takes the values $\{10^{-2}, 10^{-1.5}, 10^{-1}, 10^{-0.5}, 10^0, 10^{0.5}\}$. This corresponds to the memory factor $\alpha = e^{-\beta}$ taking the values $\{0.99, 0.96, 0.90, 0.73, 0.37, 0.042\}$. These plots demonstrate the important fact that the performance gain is a convex function of β. On the one hand, when β is small, the user mobility has a high memory level, which favors the predictive scheme. On the other hand, at $\alpha = 0$, the variance of $\sum_{i=0}^{n} v_i$ reaches a local minimum, $n\sigma^2$. Thus, when β is large (i.e., α is small) and μ is not too small (larger than 0.1 in this case), the disadvantage of the nonpredictive scheme is mainly determined by a mobile's average velocity as explained below. Therefore, in this case, since μ does not affect the cost of the predictive scheme, the predictive performance gain is larger for larger β. In other words, a smaller variance in the mobile velocity, as observed over the location inspection periods, reflects a more fluid-flow-like movement pattern, which favors the predictive scheme.

When $\mu = 0$, the performance gain decreases from about 2.5 down to unity, as the memory factor of the system decreases from 0.99 to 0.042. In particular, for $\mu = 0$ and $\alpha \approx 0$, the mobility of the mobile has the pattern of random-walk. In this case, the predictive scheme does not have any advantage over the nonpredictive one. However, in all other cases, the predictive scheme results in substantial savings.

Fig. 5(b) presents the plots of the performance gain versus μ, for various values of σ, and with fixed $\beta = 10^{-0.5}$ ($\alpha = 0.73$). These plots demonstrate that the performance gain is a
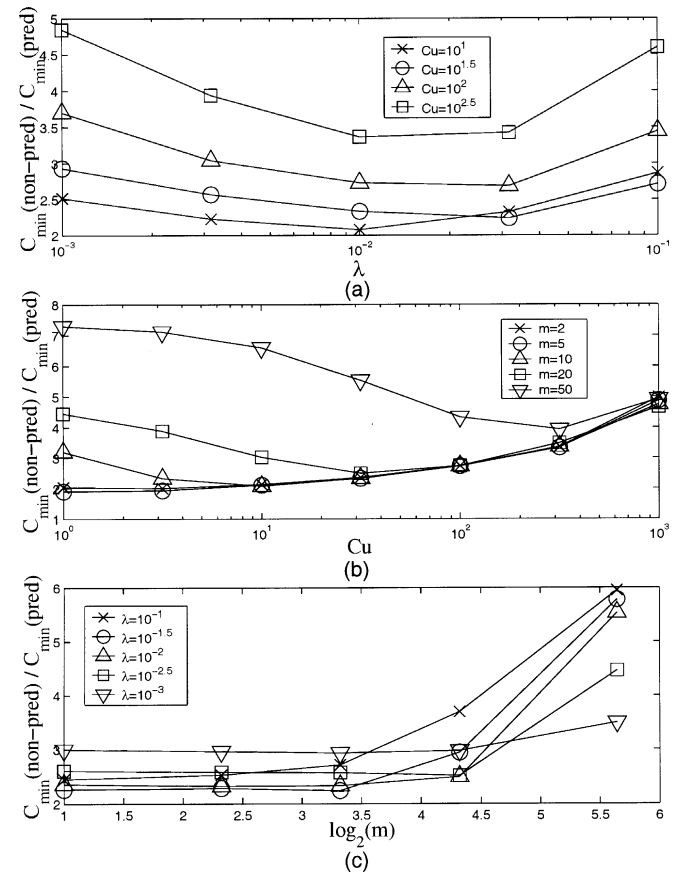
faster-than-linearly increasing function of μ. Maximum savings are achieved when $\mu \gg \sigma$, since, in this case, the mobile mobility pattern is close to the fluid-flow model, where a mobile's velocity and location are easily predictable. In this case, the mobile never needs to update its location, and therefore, the only cost incurred using the predictive scheme is the cost of paging once in the cell of the predicted location, when a call arrives, to verify that the mobile is indeed there. The performance gain can grow without bound as μ approaches infinity. For example, when $\sigma = 0.1$ and $\mu = 10$, a performance gain greater than 1000 can be achieved. However, when $\mu = 0.1$ and $\sigma > 0.1$, the performance gain is close to unity.

Fig. 5(c) shows the plots of the performance gain versus σ, for various values of β, and with fixed $\mu = 0.5$. These plots demonstrate that, when the memory level is high enough (e.g., $\alpha > 0.73$), the performance gain is a convex, nonmonotonic function of σ. On the one hand, for small values of σ, as shown in the last figure, the mobility pattern become more like the fluid-flow model and, hence, the predictive performance gain increases. On the other hand, for large σ, the advantage of using prediction becomes more prevalent, as long as the memory level is high. When the memory level is relatively low (e.g., $\alpha \leq 0.73$), the performance gain is approximately an exponentially decreasing function of σ.

In Fig. 6, we study how the parameters (λ, C_u, m) affect the performance gain. We set the mobility parameters $(\beta, \mu, \sigma) = (10^{-0.5}, 0.5, 0.5)$.

Fig. 6(a) presents the plots of the performance gain versus λ, for various values of C_u, and with fixed $m = 10$. These plots demonstrate that the predictive performance gain is a convex function of λ. On the one hand, when the call arrival rate is low, the interarrival time becomes longer, and a mobile drifts farther away from the location where the last call arrived. This leads to higher updating and paging costs for the nonpredictive scheme. On the other hand, when the call arrival rate is high, as shown in Section IV-A, the cost of the predictive scheme approaches one. In this case, the disadvantage of the nonpredictive scheme, due to the nonzero mean velocity, becomes more dominant as λ becomes larger. In particular, the transition point is around $\lambda = 0.01$ for the given set of parameters.

Fig. 6(b) shows the plots of the performance gain versus C_u, for the various values of m, and with fixed $\lambda = 0.01$. These plots demonstrate that, when m is small (e.g., $m \leq 5$), the performance gain is an approximately linearly increasing function of the updating cost to paging cost ratio. However, when m is large (e.g., $m \geq 10$), the performance gain is a convex function of C_u (note that the C_u axis is in the logarithmic scale). As a call arrives in-between location inspections, a mobile may have traveled beyond the updating distance. In the nonpredictive scheme, the system needs to page in more cells than it does in the predictive scheme. This problem of the nonpredictive scheme becomes more prominent when the paging cost to updating cost becomes larger, i.e., when C_u becomes smaller. Furthermore, the disadvantage of the nonpredictive scheme is more apparent as m increases. This is studied in more detail in the next figure.

Fig. 6(c) presents the plots of the performance gain versus m, for the various values of λ, and with fixed $C_u = 10^{1.5}$. These plots demonstrate that the performance gain is an increasing function of the location inspection period. This shows that the predictive scheme improves the performance of a system where the location inspection cost is relatively high.

C. Dynamic Gauss–Markov Parameter Estimation

We use simulation to study the performance of the proposed scheme in systems where the mobility pattern cannot be ideally represented by the Gauss–Markov model. We simulated the 2-D random way-point mobility model as described in Section II-A-3. The time-varying Gauss–Markov parameters are estimated as described in Section II-C and applied to the predictive scheme as described in Section II-B.

Fig. 7 illustrates the effectiveness of combining dynamic parameter estimation and predictive mobility management. In both figures, the parameter estimation window is $w_{\text{est}} = 10$. The call arrival rate λ takes the values $\{10^{-1}, 10^{-2}, 10^{-3}\}$. The cost per location update is assumed to be $C_u = 10[C_p]$.[5] We further assume that $m = 1$. In the random way-point model, we assume that the mobile velocity has unit magnitude with zero pause time (i.e., when a mobile reaches a destination point, it immediately moves toward a new destination). The system coverage area is assumed to be a square with side length W. Clearly, the larger the W, the more predictable is the mobile movement. In particular, when W approaches infinity, a mobile appears to move with constant velocity with a pre-selected random direction.

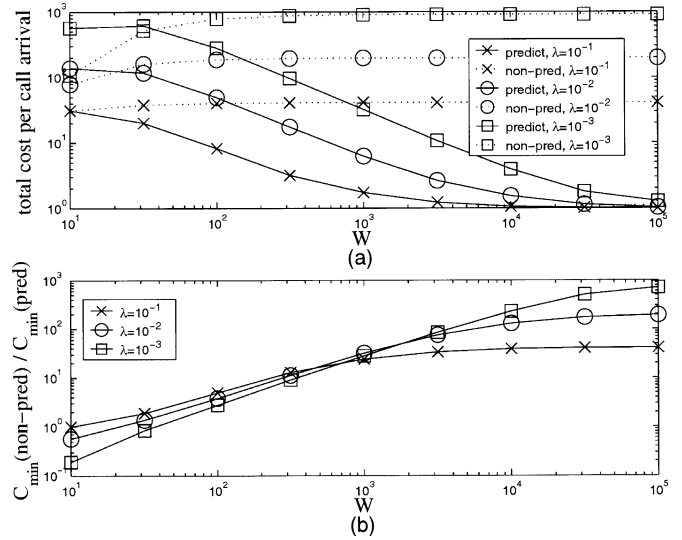[5]Simulation results with other C_u values yield similar results and are omitted here.

Fig. 7. Dynamic parameter estimation with random way-point mobility model, for $C_u = 10$. (a) Comparison of minimum (over N) cost. (b) Performance gain.

Fig. 7(a) plots the minimum total cost (obtained with the optimal location threshold) of location updating and mobile paging, per call arrival, for both the predictive scheme and the nonpredictive scheme. This figure shows that, as W increases, the mobile movement approaches the constant-velocity fluid-flow model, and the cost of the nonpredictive scheme approaches its maximum value. This maximum value can be determined analytically, since when a mobile moves with unit constant velocity,

$$C_{\text{total}}^{np}(N) = \int_0^\infty \lambda e^{-\lambda t} \left(C_u \frac{t}{N} + \frac{1}{N} \int_0^N \pi s^2 \, ds \right) dt. \quad (65)$$

It is easy to optimize N in the above and to obtain the minimum cost of the nonpredictive scheme $C_{\text{total}}^{np}(N_{\text{opt}}^{np}) = 1.92(C_u/\lambda)^{2/3}$. The simulation model can be verified by substituting the values of C_u and λ into the above equation and comparing the results with the simulation results.

In contrast, as W increases, the cost of the predictive scheme decreases. As expected, this cost approaches $1[C_p]$ when W is very large. This is the case where the mobile location is perfectly predictable, and, hence, the best scheme is not to update at all, and each time a call arrives, the system pages the mobile once.

Fig. 7(b) plots the performance gain of the predictive scheme as defined in (64). This figure shows that when W is not too large, the performance gain is approximately linearly increasing with W. When W is very large, as expected from the results shown in the previous figure, the performance gain levels off and approaches its global maximum. As a point of reference, for the moderate case where $C_u = 10$ and $\lambda = 0.01$, the predictive scheme outperforms the nonpredictive scheme by a maximal factor of 200!

It is worth noting that, in the case of dynamic Gauss–Markov parameter estimation, the predictive performance gain can degrade to less than 1 when W is very small. This is in contrast with the ideal case shown in Section IV-B, where the predictive scheme performs as well as the nonpredictive scheme even

when the mobile movement is memoryless. This performance degradation is a direct result of our suboptimal parameter estimation algorithm. For example, if we had *a priori* knowledge (e.g., through long-term observation) that the mobile movement follows the random way-point pattern, a better parameter estimation using $w_{\text{est}} = 1$ could be employed. For systems with general mobility patterns, improving the parameter estimation algorithm would further improve the performance gain of the predictive scheme beyond what is shown in Fig. 7(b).

V. CONCLUSION

Mobile users in PCS networks move with wide variety of mobility patterns, especially in networks with multilayer macrocellular and microcellular infrastructures [10]. We have presented a novel predictive distance-based mobility management scheme, which takes full advantage of the correlation between a mobile's current velocity and location and its future velocity and location. We have introduced a mobility model based on the Gauss–Markov random process, which captures the various degrees of velocity correlation in time. It even includes both the random-walk and fluid-flow models as its two extreme cases. An analytical framework is introduced to evaluate the performance of the proposed scheme. The analytical framework enables us to study the effects of various parameters on the mobility management cost, to optimize the location inspection frequency and location updating threshold through numerical analysis, and to determine the relative performance of the proposed scheme over the nonpredictive scheme.

The numerical results have demonstrated the important fact that the performance gain of the predictive scheme is not an increasing function of the memory level of a mobile's movement. This validates our observation that the *memory level* does not directly indicate the *predictability* of a mobile's movement. Overall, however, the numerical analysis confirms our intuition that the cost of predictive mobility management is inversely proportional to the predictability of a mobile's mobility pattern. In particular, our numerical results suggest that this cost is a monotonic function of the velocity variance and the call arrival rate, that it is a concave function of the memory level, and that it is independent of the mean velocity. Correspondingly, the optimal location inspection period and the optimal updating distance are monotonic functions of the velocity variance and the call arrival rate, but they are, respectively, convex and concave functions of the memory level. In the span of parameter values under consideration, the performance improvement of the predictive scheme ranges from unity, in the random-walk case, to a factor of a few orders of magnitude, in the constant velocity case. As a point of reference, as shown in Fig. 5(b), in the near fluid-flow case, where $\sigma = 0.1$, the performance gain quickly increases from 8, when $\mu = 0.3$, to 100, when $\mu = 3$. In more general terms, the mobility management cost can be reduced by more than 50% for all systems where the mobiles have moderate mean velocity (e.g., $\mu \geq 0.5$), and where performing a location update by a mobile is as least as expensive as paging a mobile in a cell (e.g., $C_u \geq 1$).

Our analysis using ideal Gauss–Markov mobility pattern has demonstrated the potentially enormous advantage of using prediction in mobility management. For practical systems with time-varying mobility patterns, we have introduced a dynamic parameter estimation algorithm to be employed in conjunction with the proposed predictive scheme. Simulation results of systems with the random way-point mobility model have demonstrated that the proposed predictive scheme can largely outperform the nonpredictive scheme, even with a suboptimal parameter estimator.

The vast performance gain of predictive mobility management, as shown by both analysis and simulation, offsets the cost of enabling prediction in the mobile terminal and the wireline backbone. This justifies the investment of such software and hardware devices in the next generation wireless personal communication networks.

APPENDIX A
DERIVATION OF C_{inspect} WITH EXPONENTIALLY DISTRIBUTED CALL INTERARRIVAL TIME

In this case, the PDF of the call interarrival time is $f_{\text{call}}(n) = \lambda(1 - \lambda)^n$. From (17), we have

$$C_{\text{inspect}} = C_i \sum_{n=0}^{\infty} \lambda(1-\lambda)^n \lfloor n/m \rfloor$$
$$= C_i \sum_{i=0}^{\infty} \sum_{j=im}^{(i+1)m-1} \lambda(1-\lambda)^j \lfloor j/m \rfloor$$
$$= C_i \sum_{i=0}^{\infty} i \sum_{j=im}^{(i+1)m-1} \lambda(1-\lambda)^j$$
$$= C_i \left(1 - (1-\lambda)^m\right) \sum_{i=0}^{\infty} i(1-\lambda)^{im}$$
$$= C_i \frac{(1-\lambda)^m}{1 - (1-\lambda)^m}. \tag{66}$$

APPENDIX B
DERIVATION OF $f_{y_0|(u_0, u_m')}(y, \xi_1, \xi_2)$

Since $u_0, u_1, \ldots, u_{m-1}, u_m'$ are jointly Gaussian, $y_0 = \sum_{i=0}^{m-1} u_i$ is jointly Gaussian along with $u_0, u_1, \ldots, u_{m-1}, u_m'$. With slight abuse of notation, we have

$$f_{y_0|(u_0, u_m')}(y, \xi_1, \xi_2)$$
$$= \frac{f_{(y_0, u_m')|u_0}(y, \xi_1, \xi_2)}{f_{u_m'|u_0}(\xi_2, \xi_1)}$$
$$= f_{(y_0|u_0)|(u_m'|u_0)}(y, \xi_1, \xi_2)$$
$$= \frac{1}{\sigma_{y_0|u_0} \sqrt{2\pi(1-\gamma^2)}}$$
$$\cdot \exp\left\{ -\frac{\left[(y - \mu_{y_0|u_0}) - \gamma \frac{\sigma_{y_0|u_0}}{\sigma_{u_m'|u_0}}(\xi_2 - \mu_{u_m'|u_0})\right]^2}{2\sigma_{y_0|u_0}^2(1-\gamma^2)} \right\} \tag{67}$$

where $\mu_{u'_m|u_0}$ and $\mu_{y_0|u_0}$ are the means of $u'_m|u_0$ and $y_0|u_0$, respectively, $\sigma_{u'_m|u_0}$ and $\sigma_{y_0|u_0}$ are the standard deviations of $u'_m|u_0$ and $y_0|u_0$, respectively, and γ is the correlation factor of $y_0|u_0$ and $u'_m|u_0$.

From the definition of u'_m, we have

$$\mu_{u'_m|u_0} = \alpha^m u_0 \tag{68}$$

and

$$\sigma^2_{u'_m|u_0} = \sigma^2(1-\alpha^2) \sum_{i=0}^{m-1} \alpha^{2(m-i-1)}$$
$$= (1-\alpha^{2m})\sigma^2. \tag{69}$$

The mean of y_0, given u_0, is

$$\mu_{y_0|u_0} = \sum_{i=0}^{m-1} \mu_{u_i|u_0} = \sum_{i=0}^{m-1} \alpha^i u_0 = \frac{1-\alpha^m}{1-\alpha} u_0. \tag{70}$$

To find a formula for $\sigma_{y_0|u_0}$, we first consider the covariance between $u_i|u_0$ and $u_j|u_0$. For any i and j satisfying $0 \le i < j \le m-1$, we have

$$\mathrm{E}[(u_i|u_0 - \mu_{u_i|u_0})(u_j|u_0 - \mu_{u_i|u_0})]$$
$$= \mathrm{E}\left[\sigma^2(1-\alpha^2) \sum_{l=0}^{i-1} \alpha^{i-l-1} w_l \sum_{l=0}^{j-1} \alpha^{j-l-1} w_l \right]$$
$$= \sigma^2(1-\alpha^2) \sum_{l=0}^{i-1} \alpha^{i-l-1} \alpha^{j-l-1}$$
$$= \sigma^2(1-\alpha^{2i}) \alpha^{j-i}. \tag{71}$$

Then

$$\sigma^2_{y_0|u_0} = \mathrm{E}\left[\left(\sum_{i=0}^{m-1} u_i|u_0 - \mu_{u_i|u_0} \right)^2 \right]$$
$$= \sum_{i=0}^{m-1} \mathrm{E}\left[(u_i|u_0 - \mu_{u_i|u_0})^2 \right] + 2 \sum_{i=0}^{m-1} \sum_{j=i+1}^{m-1}$$
$$\cdot \mathrm{E}\left[(u_i|u_0 - \mu_{u_i|u_0})(u_j|u_0 - \mu_{u_j|u_0}) \right]$$
$$= \sum_{i=0}^{m-1} \sigma^2_{u_m|u_0} + 2 \sum_{i=0}^{m-1} \sum_{j=i+1}^{m-1} \sigma^2(1-\alpha^{2i})\alpha^{j-i}$$
$$= \frac{-(1-\alpha^m)^2 + m(1-\alpha^2) - 2\alpha(1-\alpha^m)}{(1-\alpha)^2} \sigma^2. \tag{72}$$

Finally, the covariance factor between u'_m and y_0 is

$$\gamma = \frac{C_{(u'_m, y_0)|u_0}}{\sigma^2_{u'_m|u_0} \sigma^2_{y_0|u_0}} \tag{73}$$

where

$$C_{(u'_m, y_0)|u_0}$$
$$= \mathrm{E}\left[(u'_m|u_0 - \mu_{u'_m|u_0}) \left(\sum_{i=0}^{m-1} u_i|u_0 - \mu_{u_i|u_0} \right) \right]$$
$$= \sum_{i=0}^{m-1} \mathrm{E}\left[(u'_m|u_0 - \mu_{u'_m|u_0})(u_i|u_0 - \mu_{u_i|u_0}) \right]$$
$$= \sum_{i=0}^{m-1} \sigma^2(1-\alpha^{2i})\alpha^{m-i}$$
$$= \frac{(1-\alpha^m)(\alpha-\alpha^m)}{1-\alpha} \sigma^2. \tag{74}$$

REFERENCES

[1] "Cellular radio-telecommunications intersystem operation: Automatic roaming," EIA/TIA, Tech. Rep. IS-41.3-B, 1991.
[2] I. F. Akyildiz and J. S. M. Ho, "Dynamic mobile user location update for wireless PCS networks," *Wireless Netw.*, vol. 1, no. 2, pp. 187–196, 1995.
[3] I. F. Akyildiz, J. S. M. Ho, and Y.-B. Lin, "Movement-based location update and selective paging for PCS networks," *IEEE/ACM Trans. Networking*, vol. 4, pp. 629–638, Aug. 1996.
[4] I. F. Akyildiz *et al.*, "Mobility management in next-generation wireless systems," *Proc. IEEE*, vol. 87, pp. 1347–1384, Aug. 1999.
[5] A. Bar-Noy, I. Kessler, and M. Sidi, "Mobile users: To update or not to update?," *Wireless Netw.*, vol. 1, no. 2, pp. 175–185, 1995.
[6] A. Bhattacharya and S. K. Das, "LeZi-update: An information-theoretic approach to track mobile users in PCS networks," in *Proc. ACM MOBICOM*, Seattle, WA, Aug. 1999, pp. 1–12.
[7] J. Broch, D. A. Maltz, D. B. Johnson, Y.-C. Hu, and J. Jetcheva, "A performance comparison of multi-hop wireless ad hoc network routing protocols," in *Proc. ACM MOBICOM*, Dallas, TX, Oct. 1998, pp. 85–97.
[8] T. X. Brown and S. Mohan, "Mobility management for personal communication systems," *IEEE Trans. Veh. Technol.*, vol. 46, pp. 269–278, May 1997.
[9] S. R. Das, C. E. Perkins, and E. M. Royer, "Performance comparison of two on-demand routing protocols for ad hoc networks," in *Proc. IEEE INFOCOM*, Tel Aviv, Israel, Mar. 2000, pp. 3–12.
[10] A. Ganz, Z. J. Haas, C. M. Krishna, and D. Tang, "On optimal design of multi-tier wireless cellular systems," *IEEE Commun. Mag.*, vol. 35, pp. 88–93, Feb. 1997.
[11] A. Gelb, *Applied Optimal Estimation*. Cambridge, MA: MIT Press, 1974.
[12] Z. J. Haas and Y.-B. Lin, "On optimizing the location update costs in presence of database failures," *Wireless Netw.*, vol. 4, no. 5, pp. 419–426, 1998.
[13] M. Hellebrandt, R. Mathar, and M. Scheibenbogen, "Estimating position and velocity of mobiles in a cellular radio network," *IEEE Trans. Veh. Technol.*, vol. 46, pp. 65–71, Feb. 1997.
[14] J. S. M. Ho and I. F. Akyildiz, "Mobile user location update and paging under delay constraints," *Wireless Netw.*, vol. 1, no. 4, pp. 413–425, 1995.
[15] L.-R. Hu and S. S. Rappaport, "Adaptive location management scheme for global personal communications," *Inst. Elec. Eng. Proc.—Commun.*, vol. 144, no. 1, pp. 54–60, Feb. 1997.
[16] Y.-C. Hu and D. B. Johnson, "Caching strategies in on-demand routing protocols for wireless Ad Hoc network," in *Proc. ACM MOBICOM*, Boston, MA, Aug. 2000, pp. 231–242.
[17] K.-R. Koch, *Parameter Estimation and Hypothesis Testing in Linear Models*. New York: Springer-Verlag, 1999.
[18] Z. Lei and C. Rose, "Wireless subscriber mobility management using adaptive individual location areas for PCS systems," in *Proc. IEEE Int. Conf. Communications*, vol. 3, June 1998, pp. 1390–1394.
[19] J. Li, H. Kameda, and K. Li, "Optimal dynamic mobility management for PCS networks," *IEEE/ACM Trans. Networking*, vol. 8, pp. 319–327, June 2000.

[20] B. Liang and Z. J. Haas, "Predictive distance-based mobility management for PCS networks," in *Proc. IEEE INFOCOM*, New York, NY, Mar. 1999, pp. 1377–1384.

[21] U. Madhow, M. L. Honig, and K. Steiglitz, "Optimization of wireless resources for personal communications mobility tracking," *IEEE/ACM Trans. Networking*, vol. 3, pp. 698–707, Dec. 1995.

[22] S. Mohan and R. Jain, "Two user location strategies for personal communications services," *IEEE Pers. Commun.*, vol. 1, no. 1, pp. 42–50, 1st Qtr. 1994.

[23] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. New York: McGraw-Hill, 1991.

[24] C. Rose and R. Yates, "Minimizing the average cost of paging under delay constraints," *Wireless Netw.*, vol. 1, no. 2, pp. 211–219, 1995.

[25] C. Rose, "Minimizing the average cost of paging and registration: A timer-based method," *Wireless Netw.*, vol. 2, no. 2, pp. 109–116, 1996.

[26] C. Rose, "State-based paging/registration: A greedy technique," *IEEE Trans. Veh. Technol.*, vol. 48, pp. 166–173, Jan. 1999.

[27] D. Tang and M. Baker, "Analysis of a metropolitan-area wireless network," in *Proc. ACM MOBICOM*, Seattle, WA, Aug. 1999, pp. 13–23.

[28] D. Tang and M. Baker, "Analysis of a local-area wireless network," in *Proc. ACM MOBICOM*, Boston, MA, Aug. 2000, pp. 1–10.

[29] E. Wong and B. Hajek, *Stochastic Processes in Engineering Systems*. New York: Springer-Verlag, 1985.

[30] V. W. S. Wong and V. C. M. Leung, "Location management for next-generation personal communications networks," *IEEE Network*, vol. 14, pp. 18–24, Sept./Oct. 2000.

[31] V. W. S. Wong and V. C. M. Leung, "An adaptive distance-based location update algorithm for next-generation PCS networks," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 1942–1952, Oct. 2001.

[32] H. Xie, S. Tabbane, and D. J. Goodman, "Dynamic location area management and performance analysis," in *Proc. IEEE Vehicular Technology Conf.*, Secaucus, NJ, May 1993, pp. 536–539.

[33] H. Xie and D. J. Goodman, "Mobility models and biased sampling problem," in *Proc. IEEE Int. Conf. Universal Personal Communications*, vol. 2, Ottawa, ON, Canada, Oct. 1993, pp. 803–807.

Ben Liang (S'94–M'01) received the B.Sc. and M.Sc. degrees in electrical engineering from Polytechnic University, Brooklyn, NY, in 1997 and the Ph.D. degree in electrical engineering from Cornell University, Ithaca, NY, in 2001.

During the 2001–2002 academic year, he was a Lecturer and Postdoctoral Associate at Cornell University. He joined the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada, in 2002.

Zygmunt J. Haas (S'84–M'88–SM'90) received the B.Sc. degree in electrical engineering in 1979 and the M.Sc. degree in electrical engineering in 1985. He received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1988.

He joined the Network Research Department, AT&T Bell Laboratories in 1988, where he worked on wireless communications, mobility management, fast protocols, optical networks, and optical switching. From September 1994 to July 1995, he was with the AT&T Wireless Center of Excellence, where he investigated various aspects of wireless and mobile networking, concentrating on TCP/IP networks. In August 1995, he joined the faculty of the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY. He is an author of numerous technical papers and holds 15 patents in the fields of high-speed networking, wireless networks, and optical switching. His research interests include mobile and wireless communication and networks, personal communication service, and high-speed communication and protocols.

Dr. Haas has organized several workshops, delivered tutorials at major IEEE and ACM conferences, and serves as editor of several journals and magazines, including the IEEE TRANSACTIONS ON NETWORKING and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He has been a guest editor of several issues of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. He is a member of the Association for Computing Machinery and the Chair of the IEEE Technical Committee on Personal Communications.