The Power of Mobility Prediction in Reducing Idle-State Signalling in Cellular Systems: A Revisit to 4G Mobility Management

Seyed Amir Hoseinitabatabei, *Member, IEEE,*
Abdelrahim Mohamed, *Member, IEEE,* Masoud Hassanpour, *Member, IEEE,*
and Rahim Tafazolli, *Senior Fellow, IEEE*

Abstract

Conventional mobility management schemes tend to hit the core network with increased signaling load when the cell size is shrinking and the user mobility speed increases. To mitigate this problem research community has proposed various intelligent mobility management schemes that take advantage of the predictability of the users mobility pattern. However, most of the proposed solutions are only focused on signaling of the active-state (i.e., handover signaling) and proposals on improvement of the idle-state signaling has been limited and were not well received from the industrial practitioners. This paper first surveys the major shortcomings of the existing proposals for the idle mode mobility management and then proposes a new architecture, namely predictive mobility management (PrMM) to mitigate the identified challenges. An analytical framework is developed and a closed form solution for the expected signaling overhead of the PrMM is presented. The results of numerical evaluations confirm that, depending on user mobility and network configuration, the PrMM efficiency can surpass the long term evolution (LTE) 4G signaling scheme by over 90%. Analysis of the results shows that the best performance is achieved at highly dense paging areas and lower cell crossing rates.

Index Terms

Mobile communication networks; mobility management; predictive models, signaling

Seyed Amir Hoseinitabatabei, Abdelrahim Mohamed, Masoud Hassanpour, and Rahim Tafazolli are with 5G Innovation Center at the Institute for Communication Systems (ICS) of the University of Surrey, Surrey,United Kingdom.
E-mail: a.taba; abdelrahim.mohamed; m.hassanpour; r.tafazolli @surrey.ac.uk

I. Introduction

Wireless mobile communication systems undertake various mechanisms to locate the inactive users to deliver the incoming calls."Location update" and "paging" are the common mechanisms for tracking the idle mobile users and are grouped under the umbrella term of mobility management. A low-overhead yet fast and robust mobility management is a critical requirement for mobile network operators. This is down to the fact that some of the key performance indicator (KPI) parameters, such as connection establishment time, are strongly dependent on performance of the mobility management mechanism. Although the state-of-the-art solutions fulfill the requirements for prompt delivery of the incoming services, they hit the core with a significant signaling load when the cell size is reduced and user mobility speed increases.

Analysis of the field data from various long term evolution (LTE) networks deployments in the urban area shows that the signaling associated with idle mobility management procedures, accounts for more than 30% of the overall signaling load at the mobility management entity (MME) [1]. The signaling problem tends to intensify with the introduction of the small cells [2]. The standard mobility management schemes are based on the fundamental assumption that the user equipment (UE) moves in a random manner. However, recent studies on users mobility such as [3] and [4] suggest that in contrast to the assumption above, users locations are highly correlated and predictable. Such observations have sparked a new wave of research on new mobility management solutions that reduce the signaling overhead by predicting the UE location.

Majority of the predictive mobility management proposals concerns only the UE active-state or handover signaling. In addition, the limited works that were proposed for the idle-state were not well received from the industrial practitioners. Reasons include the excessive computation load that the alternative approaches add to the mobility management node and the undesirable effect of prediction failures on the KPIs. Moreover, the existing models do not provide solutions for the excessive UE-network signaling when predictive algorithms are in the training phase. Finally, some critical questions about the practical implications of these approaches have remained unanswered. For instance, it is not clear how user behavior and network configurations outside their simulation constraints can affect the signaling overhead of the proposed approaches.

In this study, we present a novel architecture for predictive mobility management (PrMM) in idle-state as an initial step towards mitigating the shortcomings of the existing solutions. The major design considerations for PrMM include: minimization of the signaling over both core-network and air interface, support for distributed computing for computation intensive tasks, and considering the standard approach (i.e., standard MME procedure in LTE) as an upper bound for the connection establishment time. The proposed predictive method not only reduces the expected signaling cost but guarantees that the delay for the predictive UE localization never exceeds the delay of the standard procedures. Moreover, the proposed solution utilizes the computation resources of UE and performs the required processes in a distributed manner to reduce the computation load on the network whilst avoiding any extra signaling between UE and the network during the training phase. It is worth mentioning that the final decision can be taken by either the UE or the network depending on the employed mobility management policy.

We provide an analytical framework to study performance of the proposed solution against the standard LTE scheme at different user mobility modes and network configurations. The analytical framework is flexible to accommodate different network configurations. Also, our formulation is independent of the mobility model of the user. In short, we make the following contributions.

- We introduce a novel idle-mode PrMM architecture for mobile communication networks that utilizes the repetitive patterns in UE mobility to reduce the signaling load.
- We provide a new formulation for the signaling cost in conventional mobility management based on UE mobility characteristics, paging probability and network arrangements.
- We provide an analytical framework to formulate the PrMM signaling cost as a function of the predictability of the user mobility patterns, paging probability and network characteristics. To achieve realistic estimations, a systematic compression algorithm for communication of the predicted position between the UE and MME is developed as a part of the solution.
- We study performance of the proposed solution through a series of numerical evaluations. We report on sensitivity analysis of the proposed solution on both network and user behavior parameters and make recommendations on real-world applications of the proposed solution.

The remainder of this article is organized as follows. Section II provides an overview of LTE mobility management. Section III surveys the state-of-the-art. We introduce the PrMM and elaborate on features that sets PrMM apart from state-of-the-art solutions in Section IV. Section V introduces the principals of our analytical framework. Section@VI focuses on the formulation of the conventional LTE mobility management signaling procedures. Section VII elaborates on the formulation of the signaling cost of the PrMM. Section VIII presents the evaluation results and discussions. Concluding remarks and suggestions for future research are provided in Section IX.

II. LTE Mobility Management

The provided deliberations in this section are according to the 3GPP specifications of the mobility management in [5] and [6]. As indicated earlier, MME handles the mobility management in LTE. The terms MME and mobility management unit are used interchangeably hereafter. MME is connected to a number of Evolved Node Base stations (eNB) via S1 links as is shown in fig. 1. The radio coverage of the eNBs is known as a cell, and each cell has a unique cell ID.

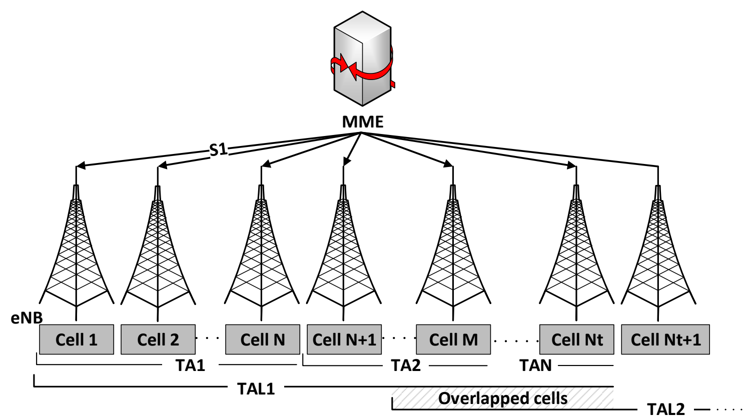


Fig. 1: MME, eNB, TA and TAL arrangements in LTE architecture.

The cells are grouped into tracking areas (TAs). Cell to TA assignment is performed during the network planning phase and is typically fixed in the operation phase. TAs also have unique IDs which are referred to as (TAI). The latter is periodically broadcasted in the coverage area. TAs are further grouped into tracking area lists (TALs). Different from TAs, size and shape of the TALs are customized for each UE. TALs can overlap with each other over a number of cells. MME provides the UE with the TAL that contains the TA in which it resides. In this study, similar to [7], TAL assignment follows the central policy. According to central policy, the cell where the UE resides belongs to a TA that is the central TA of the TAL that is assigned to the UE. Hence, our proposed modeling does not optimise the TA but rather it investigates gains of mobility prediction under existing TA policies.

In the idle-mode, the UE location is provided to the network through the tracking area updates (TAUs). While moving between the cells, if UE enters a TA that is not listed in its TAL, it requests a new TAL from the MME through the TAU procedure. Once TAU procedure is completed the TAU timer (i.e., T3412) is set which upon expiry forces the UE to perform another TAU. The localization accuracy in idle-mode is limited to the current TAL of the UE.

Upon reception of a downlink data notification, the MME initiates the paging procedure to locate the UE and starts a timer (i.e., T3413) to wait for the responses. S1-AP (S1-application protocol) paging messages are sent to all eNBs within the TAL of the UE. Each eNB that receives the paging request decodes the message to obtain the ID of the UE and creates a corresponding paging message that is to be broadcasted, at the next paging occasion, to all UE within its coverage area. eNBs that receive the paging message from MME make up the paging area (PA). It should be noted that the size of the PA has a direct impact on the generated signaling load. UE in idle mode checks for paging once every discontinuous reception (DRX) cycle. The UE checks the UE identity in the broadcasted messages. If the UE find its identity, it will trigger random access procedure to establish RRC connection and the eNB responds with RRC connection setup message to establish the connection.

As a final remark we would like to emphasis on the trade-off between the signaling costs of TAU and paging on the TAs size. While with increasing the number of cells within each TA the frequency of TAUs decreases, the paging overhead increases due to the presence of more cells in the PA. We will return to this effect when analyzing our simulation results in Section VIII.

III. Related Work

Existing works on idle-mode mobility management can be categorized into three groups. These include sequential paging, methods for Reconfiguration of Network Parameters (RNP), and distributed mobility management (DMM) schemes. Figure 2 represents a taxonomy of the existing approaches.

Sequential paging techniques strive to minimize the paging cost by multi-step paging of a number of PAs in the order of the likelihood of the user presence. Such techniques face the challenge of how to determine the optimum size of the PAs at each step. Several solutions were proposed to address the aforementioned challenge. These can be grouped into deterministic methods and probabilistic methods. The former define common PAs for all the users at each phase of the paging process, e.g., [7], [8]. Such an approach is effective in reducing the paging load on the radio interface but may add a significant delay to the connection establishment time. On the other hand, the probabilistic (or smart) paging solutions, such as [9], [10] aim to create personalized PAs through learning and predicting the user mobility pattern at the MME. In comparison with deterministic methods, probabilistic solutions impose more computation overhead. Still, due to the personalized PA assignment they are more effective in reducing

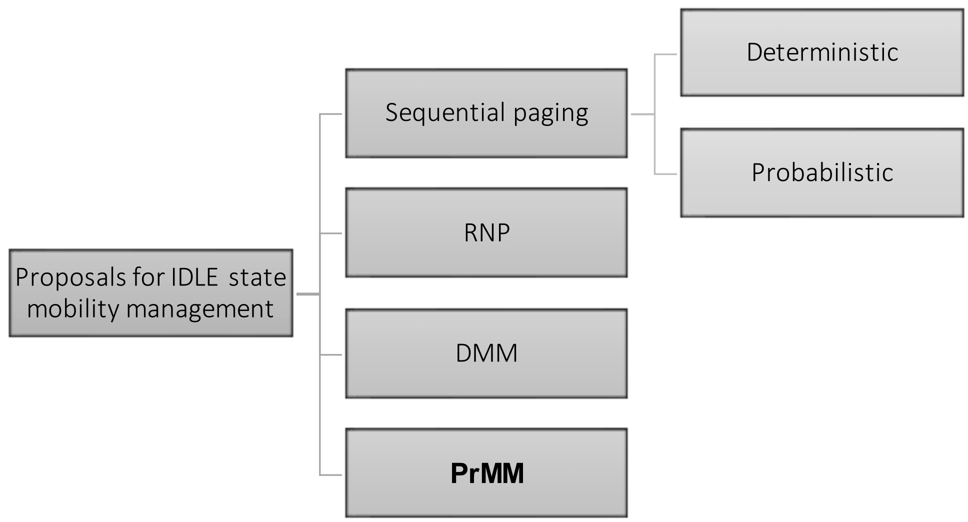


Fig. 2: Summary of the existing approaches for the mobility management in Idle-mode.

the paging cost. probabilistic (Smart) paging solutions are in general prone to the following problems. Firstly, the communication load between the mobility management node and the UE increases during the training phase. The second issue is the limited scalability due to the computation burden and memory requirements for training and maintaining an individual model for each user. Thirdly, if the predictions fails, the paging procedure has to initiate further paging messages. This as a result increases the connection establishment time. Finally, the realization of the solutions above requires new mechanisms to be created for the storage and management of the user mobility context information within the network. It is worth mentioning that literature on probabilistic mobility management includes predictive handover schemes e.g., [11], [12]. However, the reader should note that these studies are related to the mobility management in active mode and in this regard fall outside the scope of this study.

The RNP techniques consider the trade-off between the cost of paging and TAU in trying to find an optimum arrangement of the cells in the TAs. They are more costly and complicated in execution as compared with Sequential Paging methods. Proposals in this are include the work of [13] and [14]. On the other hand, the DMM approaches improve the mobility management traffic by deploying distributed mobility anchor points close to terminal locations. It follows the recent paradigm change towards a flat and distributed architecture for next generation mobile networks [15]. DMM assumes that instead of MME, the mobility management functions are performed by multiple Mobility Agents (MA) (i.e. equivalent to eNB in LTE architecture). The main challenge ahead of DMM is to locate the UE from the distributed location databases at

MAs. Existing solutions includes using distributed hash tables and Bloom filters [16].

In addition to mobility management and optimization of the associated signalling, context information (in particular user location) have also been widely investigated to optimize network performance and operating parameters such as transmission mode and layers, scheduling optimization, interference management, handover timers, cell range extension/offset parameters, and location-based services. Such information can also be used to maximize network coverage/capacity or to minimize network resources. The scheme patented in [17] utilizes previous user location to predict future locations at the network side in order to optimize the network parameters. The authors of [18] and [19] exploit user location to identify good and bad coverage areas to optimize network coverage. In [20], big data including user location along with other network and user parameters are used to improve the user quality of experience.

IV. Proposed Predictive Mobility Management: PrMM

Herein, we introduce a new architecture for the mobility management in idle mode. Our proposed approach takes advantage of the predictions on repetitive patterns of the user mobility and in this regard is entitled as Predictive Mobility Management (i.e. PrMM). In this section we explain how PrMM improves on the shortcomings of the state of the art solutions. It is worth mentioning that the proposed solution, PrMM is equally applicable to GSM, GPRS, UMTS and LTE networks. However, to harmonize the notations and ease the elaboration of the new functions, LTE is used as the reference network architecture hereafter. The major design considerations for PrMM include: minimization of the signaling over both core network and the air interface, support for distributed computing for computation intensive tasks, and considering delay with the standard approach i.e. EPS Mobility Management (or EMM) as an upper bound for the connection establishment time.

The proposed PrMM comprises two software agents; one is at UE, and the other one resides at the MME. Inside UE, PrMM runs a data acquisition application service which collects UE location alongside with other context information e.g. time, date. The location data could be provided in the form of nearby Cell-IDs or latitude and longitude (from GPS or a hybrid approach with inertial sensors and GPS such as [21]). While Cell-IDs are available to UE at almost no extra cost, use of GPS and INS-based solutions are prone to the added computation and energy costs. In, [21] we have shown that a typical GPS on smartphones consumes up to 430mW, while INS consumption is about 0.75 of GPS and the added computational overhead is negligible.

However INS is prone to drifting error on long runs. A hybrid solution that comprises both INS and GPS with smart management of duty cycles can be considered as an accurate and energy efficient alternative to GPS.

In this study coarse location based on Cell-IDs suffice and our analysis in the next section is independent of the positioning technology that is utilised. The UE predictive model agent predicts a sequence of UE locations at the cell level rather than the exact location. In other words, the predictive scheme exploits cell-level location information rather than the exact UE position. The MME utilizes the provided predictions to track the cell-level UE location. If location-based services are available at the UE then such information can be used for mobility management. For instance, the UE can use its exact location, direction, speed to predict future locations, or such information can be reported to the network to estimate the UE future location, e.g., based on the angular span of the UE direction/speed as in the active-mode predictive scheme of [22]. In the proposed model, we consider idle-mode mobility management. Consequently, we utilise the cell-level location information rather than the exact UE position to ensure minimal increase in power consumption in the idle-state.

In parallel with data acquisition, PrMM agent analyses the patterns within the collected data and extracts their correlation with the user location. Once sufficient statistics are provided the PrMM agent sends an update containing a sequence of most likely future locations of the UE to PrMM agent at MME. When the update message is received at MME, PrMM agent terminates the standard signaling i.e. EMM and starts the predictive mode by only listening for a new update message. From this point onward, UE will periodically monitor its location and verifies its compliance with the predicted locations in the last update message. It is worth noting that agreement between the actual location and the provided model to MME is equivalent to tracking the UE by MME but without any signaling.

In the case of any deviation from the predicted model, the UE will send a TAU message to MME, fall back into the conventional mode, and wait for another occasion when it can provide a confident prediction of the future locations. The UE keeps a copy of the predicted states that are provided to the network. During the idle state, the UE continuously check the exact cell-level location against the list of predictions to verify their validity. It is worth mentioning that the cell-level location is monitored by the UE in current standards, e.g., 4G, for cell re-selection. The cell-ID can be obtained by the downlink synchronization signals such as the primary synchronization signal (PSS) and the secondary synchronization signal (SSS) in LTE.

In other words, the UE utilises the conventional cell re-selection measurements along with the predicted values to detect when the prediction fails in order to revert back to the conventional EMM approach. Falling back to EMM mode in this situation guarantees that the overall paging delay would never exceed the delay of the standard procedure. In comparison with the state of the art approaches, PrMM provides the following advantages.

- The proposed PrMM scheme allows the UE to predict several cell-level locations in advance. These predicted cell-level locations are shared with the eNB only once, and updated when prediction failure is detected. As a result, the frequency of UE-eNB message exchange depends on success/failure of the prediction rather than the cell-crossing rate.
- In PrMM architecture, the data acquisition for training the prediction model is performed at the UE side. In this regard, no further communication between UE and network is required during the training phase. This design consideration also makes the mobility management scalable with no compromise on the convergence time of the learning algorithm.
- PrMM utilizes the existing computation resources in UEs to eradicate the computation load for user profiling and pattern analysis on MME. All the analysis are performed on the UE and the results, as a series of predicted locations, are provided to the MME.
- During the predictive mode the compliance of the UE location and the predictions are continuously monitored at the UE; therefore the network knowledge about the user location would be personal and deterministic. This is in contrast to the existing predictive models where knowledge about user location is a soft guess with a certain probability. Also, it differs from traditional deterministic models because the paging area is tailored to the mobility pattern of individual users.
- Such accurate tracking of the user location, in return, minimizes the paging load and delay as only the specific cell where UE resides would be paged.
- The suggested design is ideal for DMM, the reason is that UEs is solely responsible to store and share the mobility information with the anchor points in the network and in this regard this solution is not faced with the issues related to the management of the mobility context between MAs.

Fig. 3 summarizes the PrMM procedures. In this figure, different states of the mobility management mechanism alongside with their associated signaling messages are presented.

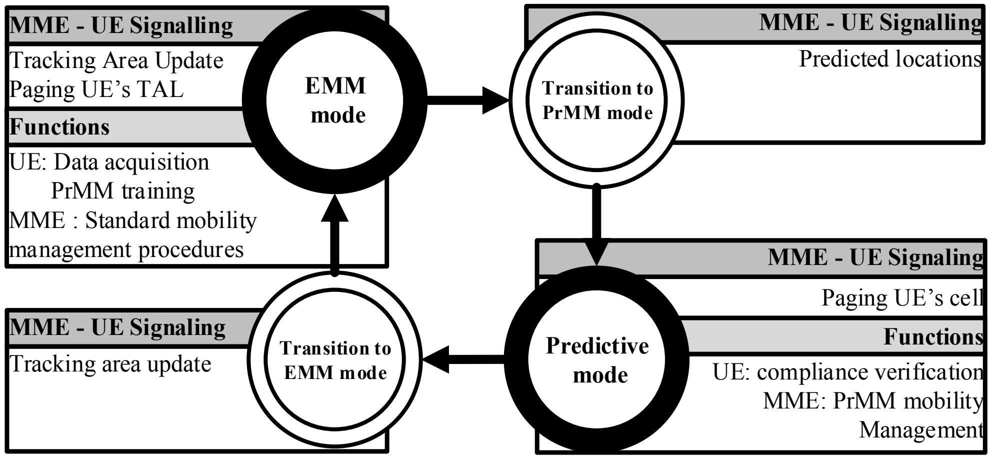


Fig. 3: Proposed predictive mobility management architecture.

V. Analytical Framework

To formulate the signaling cost, we have to make some assumption about user mobility, network configuration, and signaling overheads. We consider a random walk model for the mobility of the UE. Random walks are among the most popular mobility models for wireless mobility management studies e.g. [23], [24] and [25]. Such models are particularly desirable for predictive mobility management studies as the entropy of the user mobility can be controlled by simply adjusting the state transition probabilities. It is worth that the random walk mobility model increases the difficulty of mobility prediction, and the prediction accuracy in this model is less than deterministic models. Usually, high speed users are more predictable than low speed users, since the probability of changing the direction in the former is less than the probability in the latter. Our intention is to evaluate the lower bound of the PrMM gain, such that the gain should be higher in more deterministic mobility models. We believe that showing gains in the worst case scenario (for the predictive scheme) provides a lower bound for the performance, which in turn encourages the community to adopt such predictive approaches. To facilitate the reading, we first use a one-dimensional random walk model to formulate the UE mobility. Later we show how the 2-D framework is derived from the 1-D formulations. Our evaluations confirms that the results from 1-D mobility analysis are in compliance with 2-dimensional simulations. Similar observation was made by [7] on the study of paging and tracking area update costs.

Table I provides a list of notations and symbols used in this paper.

The idle-state period (t_c) is formulated as an exponential random variable with the mean of $\frac{1}{\lambda_c}$. Similar to [7] cell residence time (t_m) is considered to follow a gamma distribution with

TABLE I: List of notations

Symbol	Description
t_c	Idle-state period
λ_c	Rate parameter of idle-state period distribution
t_m	Cell residence time
λ_m	Rate parameter of cell residence time distribution
V	Variance of cell residence time
N_t	Number of tracking areas in TAL
N_c	Number of cells in a tracking area
c_u	Message size of tracking area update
$c_{b,p}$	Size of paging messages broadcasted from eNB to all UE
$c_{u,p}$	Size of paging messages between paged UE and MME
α	Probability of terminating the idle period with paging
N_{np}	Expected number of non-periodic tracking area updates
N_{pr}	Expected number of periodic tracking area updates
CM	Cost of conventional mode
N_{cr}	Expected number of cell crossings before UE leaves the current TAL
CPM	PrMM signalling cost
J	Total number of cells that user traverses during an idle period
C_s	Cell size
β	Prediction update overhead
$h(j)$	Correction factor
$\gamma(J)$	Likelihood of taking J steps
p	Transition probability from lower indexed state to higher indexed state
$P(j)$	Likelihood of the correct prediction of the first j steps of the user
J_{max}	Upper bound for J
p_d	Probability of horizontal movement in 2-D random walk model
p_h	Probability of forward horizontal movement in 2-D random walk model
p_v	Probability of up vertical movement in 2-D random walk model
R	Signalling reduction in PrMM w.r.t. conventional mode

the average of $\frac{1}{\lambda_m}$ and the variance of V. Fig.1 represents the arrangements of cells, tracking areas and tracking area lists. Here, each TAL comprises N_t Tracking areas whilst each tracking area includes N_c cells.

As was mentioned, the UE initiates a tracking area update procedure once entering a new TA that is not listed in its TAL. The 3GPP specification for contents of signaling messages for TAU procedure [6] designates the message contents as either compulsory or optional. To provide a realistic estimation of the paging overhead, we added all the optional components but limited them to their minimum allowable size. Also, an extra 15% overhead is added to cover the retries and errors in message transportation over the Evolved Universal Terrestrial Radio Access (EUTRA). The total size of all the messages that are communicated over the air for the tracking area update is represented by c_u and is set equal to 1462 bits.

In the case of paging procedure, the size of the messages is extracted from a 3GPP technical report [26]. Furthermore, similar to the tracking area update procedure, 15% excess overhead is added for EUTRA transport efficiency. Amongst the paging signaling messages, some are broadcasted from eNB to all UEs while others are exchanged between the paged UE and MME. Henceforth, the first category of messages are denoted as $c_{b,p}$ and set to 146 bits, and the second category as $c_{u,p}$ set to 614 bits. The message sizes are in the same order as the reported values in commercial LTE simulation environments such as WireShark.

VI. EMM SIGNALING FORMULATION

As indicated earlier, if the prediction of PrMM fails while UE is in the idle mode, MME falls back into the conventional signaling procedure of EMM. The cost of conventional mode, hereafter CM, is defined as the sum of data bits that are communicated over EUTRA for the paging and tracking area updates for one user during one idle mode interval.

Equation (1) shows the expected value of the CM that is calculated for a user. The parameters within the first parenthesis account for the cost of the paging procedure and the second parenthesis represents the cost of tracking area updates.

$$\mathbb{E}(CM) = \alpha(N_c * N_t * c_{b,p} + c_{u,p}) + (N_{np} + N_{pr}) * c_u \tag{1}$$

Here α is the probability of terminating the idle period with paging where the model starts with a UE in idle-state. N_{np} and N_{pr} in (1) are the expected number of non-periodic and periodic tracking area updates. N_{np} is calculated as the ratio between the expected numbers of cell crossings to the expected number of cell crossings required before UE enters a new cell that is not listed in its TAL. According to [7], N_{np} and N_{pr} are calculated as follows :

$$N_{pr} = \left\lfloor \frac{1}{t_m * \lambda_c} \right\rfloor \tag{2}$$

$$N_{np} = \frac{\lambda_m}{\lambda_c * N_{cr}} \tag{3}$$

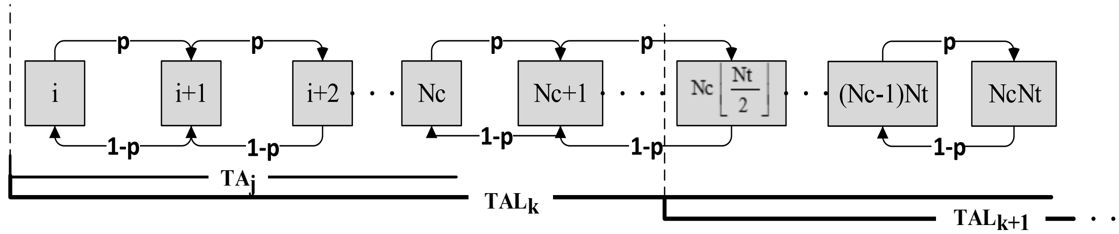


Fig. 4: Configuration of the 1-D analytical framework.

N_{cr} is the expected number of cell crossings before UE leaves the current TAL. N_{cr} is a function of user mobility characteristics and the network configuration parameters. According to [7], N_{cr} can be estimated as follows.

$$N_{cr} = \begin{cases} N_c * (\lfloor \frac{N_t}{2} \rfloor + 1) * (N_c * \lfloor \frac{N_t}{2} \rfloor + 1) & , p = 0.5 \\ \left(\frac{A}{1+A-B}\right)\left(\frac{(N_c * N_t + 1) * B - N_c * \lfloor \frac{N_t}{2} \rfloor - 1)}{2 * p - 1}\right) & \\ + \left(\frac{1-B}{1+A-B}\right)\left(\frac{(N_c * N_t + 1) * A - N_c * \lfloor \frac{N_t}{2} + 1 \rfloor - 1}{2 * p - 1}\right) & , p \neq 0.5 \end{cases} \tag{4}$$

Where,

$$A = \frac{1 - \left(\frac{1-p}{p}\right)^{N_c * (\lfloor \frac{N_t}{2} \rfloor + 1)}}{1 - \left(\frac{1-p}{p}\right)^{N_c * N_t + 1}} \quad , \quad B = \frac{1 - \left(\frac{1-p}{p}\right)^{N_c * \lfloor \frac{N_t}{2} \rfloor + 1}}{1 - \left(\frac{1-p}{p}\right)^{N_c * N_t + 1}} \tag{5}$$

The role of p, N_c and N_t parameters in the proposed analytical framework is represented in Fig. 4.

VII. PrMM FORMULATION

To elaborate the signaling cost of the PrMM; initially, we describe a generic scenario that illustrates the steps of the signaling procedure during the predictive mode. At each step, the related cost function is discussed and formulated. Finally, the cost functions are combined to form a complete representation of the PrMM overhead. Hereafter, the signaling cost of the PrMM is denoted as CPM.

Assume that during the idle periods the user crosses J cells on average. The predictive model agent on UE predicts the next J locations of the user (i.e., at the cell level) and uploads the predictions to the MME. MME utilizes the provided predictions to track the user location. The user starts to move and crosses a number of cells and tracking areas. Lets assume that after the j-th cell crossing the prediction fails. The UE must inform the MME with a TAU and the signaling system falls back into the conventional scheme, as was described in section IV.

Based on the scenario mentioned above, the main components of the CPM can be identified as: Prediction update overhead, signaling cost when predictions are correct, the signaling cost for TAU update when the prediction fails and the remaining mobility management signaling when the system falls back into the conventional scheme till the end of the idle period. Through the rest of the section, we analyze these costs one by one and then combine them into a unified model.

A. Prediction update overhead

As mentioned earlier, UE initiates the predictive mobility scheme by uploading the predicted future locations. Here we assume that the predictions are added as extra bits to the actual tracking area update message. Encapsulating the location information as raw latitude and longitude data makes the prediction overhead increasing linearly with the number of predicted locations. The problem intensifies when the cell crossing rate increases. To avoid this problem, we postulate that the UE utilizes a data compression method for location prediction updates. Herein, we propose a lossless data compression method, namely Location Data Compression (LDC). LDC takes advantage of the Geohash technique [27] to provide a unique geocode from each pair of latitude and longitude values.

Geohash method generates a set of deterministic hash codes (geocodes), based on the latitude and longitude values. One of the interesting characteristics of Geohashing is that through the hashing process, proximate locations receive same geocodes up to very last bits. As the distance between two places increases the difference between their geocodes propagates towards the first bit of the code word. LDC benefits from this property to reduce the size of the transmitted prediction information.

In this study, we assume that the geocodes are generated for the location of the center of the cells. Therefore the difference between each pair of subsequent geocodes is dependent on the size of the cells. LDC process comprises the following steps: 1) The geocode for the initial location is added to the prediction update message without any compression. 2) From the second location onwards, LDC generates a geocode for each raw location data. Only the part of the generated code that is different from the previous reference is added to the update message. 3) If the size of the chosen part is more than that of the last added point, the new point is marked as a new reference. Steps 2-3 is repeated until all the predicted locations are treated.

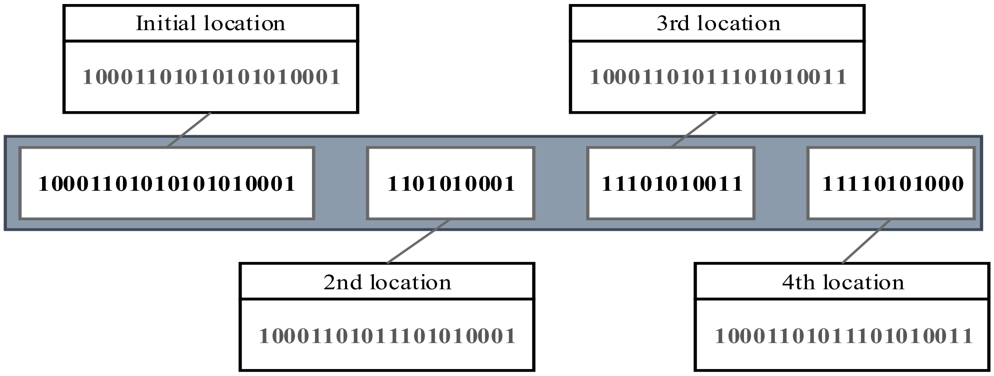


Fig. 5: An example of LDC algorithm

Fig.5 elaborates on LDC algorithm for preparing the location prediction updates. Here, the initial location information is added to the update stream in full details. The second location is different from the previous reference after the 10th bit, the third location has 11 bits of new information and finally in the fourth location the last 12 bits are distinct from the reference and are added to the stream. Geocodes of any of these locations could be used as a reference for the next coming data if the difference between their code and the new data code is shorter than their length. Once the prediction information is ready, it is encapsulated into an update message and sent to MME alongside with a tracking area update message. MME maps the predicted locations into the cell areas to track the UE. It should be noted that when locations are in the form of cell-IDs similar incremental parsing technique can be utilized to reduce the size of the prediction updates. Prediction update overhead is denoted as $\beta(J, C_s)$, where J is the total number of cells that user traverses during an idle period, and C_s is the size of the cells.

B. Cost of signaling overhead when all predictions are correct

After the initial update, while the predictions are correct, the only signaling procedure between UE and MME would be the one for paging. Since with the help of previous predictions the location of the UE for is known to the cell-level, a paging message that is initiated from the MME is forwarded only towards the cell in which the user resides. If all the predictions are correct for J steps, paging cost of the conventional scheme in (1) reduces to α * ($c_{b,p}$ +$c_{u,p}$) in the predictive mode.

In the mobility scenario that was presented at the beginning of the section, this situation is equivalent to have j = J. Here again α is the paging probability at the end of the idle period.

C. Cost of signaling overhead when some predictions are incorrect

When predictions fail, UE sends a TAU to the network and falls back into the conventional scheme, with both paging and TAU procedures. In the case of paging, all the cells within the current tracking area list of the user will be paged; therefore the cost of paging would be similar to the conventional scheme in (1) i.e. $\alpha * (N_c * N_t\, c_{b,p} + c_{u,p})$.

Also, TAU will be initiated at the end of a periodic timer or when the user leaves its current tracking area list, similar to the conventional technique. However, the idle period (t_c) is reduced by the amount of time that the user spends in the predictive mode before prediction fails. For the remainder of the idle period after the jth step of the user, λ_c in the formulation of N_{np} and N_{pr} in (2) and (3) is replaced with $h(j)$ * λ_c, where $h(j)$ is a correction factor and is calculated as follows.

$$h(j) = \frac{J - j}{J} \tag{6}$$

CPM is then calculated by integrating the introduced components as is represented in (7).

$$CPM(j, J) = \begin{cases} \alpha * (c_{b,p} + c_{u,p}) + \beta(J, C_s) * c_u & , j = J \\ (\beta(J, C_s) + 1 + N_{np}(j) + N_{pr}(j)) * c_u + \alpha * N_c * N_t * c_{b,p} + c_{u,p} & , j \leq J \end{cases} \tag{7}$$

Different from (1), here N_{np} and N_{pr} are dependent on j, that is due to the correction factor for λ_c i.e. $h(j)$. To calculate the expected value of CPM, we should sum all the possible values of CPM, weighted with their occurrence probability. According to (7), two probability functions are to be considered when calculating the expected value of CPM. First, the likelihood that user takes J steps during the t_c. Second is the likelihood that prediction fails at the j-th step.

The first probability function can be calculated based on the postulated distribution of λ_c and λ_m. According to [7] the likelihood of taking J steps is estimated by $\gamma(J)$ as:

$$\gamma(J) = \begin{cases} -\left(\frac{\lambda_c}{\lambda_m}\right)\left[1 - \left((V * \lambda_m * \lambda_c + 1)^{(\frac{-1}{V*\lambda_m^2})}\right)\right]^2 * (V * \lambda_m * \lambda_c + 1)^{(\frac{J-1}{V*\lambda_m^2})} & , J \geq 0 \\ 1 - \left(\frac{\lambda_c}{\lambda_m}\right)\left[1 - \left((V * \lambda_m \lambda_c + 1)^{(\frac{-1}{V*\lambda_m^2})}\right)\right] & , J = 0 \end{cases} \tag{8}$$

To calculate the second probability function, we need to make an assumption about the prediction accuracy or predictability of the user movements.

Entropy is probably the most fundamental quantity capturing the degree of predictability characterizing a time series. The study in [3] shows that the average entropy of the cell-level location of mobile phone subscribers is as low as 0.8. This means that when enough history of user past locations is provided the real uncertainty in a typical users whereabouts at each time is $2^{0.8} = 1.74$, i.e., fewer than two locations.

All mobility prediction models aim to capture the full spatiotemporal correlations present in a persons mobility pattern to improve their estimation of mobility entropy. Capturing the temporal correlations is typically performed by training a Markove model on user mobility history e.g. [11]. More powerful prediction approaches such as LZ algorithm [28] form an infinite order Markov chain and extract all the existing correlations in user mobility patterns to achieve an estimation of entropy that is near the actual entropy of the user mobility.

In this study we assume that the user movements are independent from each other, that is to establish a lower bound on the predictability of the user mobility. Therefore, any more realistic mobility model should improve on the reported performance of PRMM in this study. Also, we assume that the learning process of the prediction algorithm is long enough to accurately estimate the probability of moving in each directions. Since UE selects different movement directions with different probabilities, the optimum strategy for making the most accurate prediction is to always choose the most probable direction i.e.direction associated with p_m.

$$p_m = \max\{p, 1-p\} \tag{9}$$

Then, the likelihood of the correct prediction of the first j steps of the user can be formulated as follows.

$$P(j) = (p_m)^j \tag{10}$$

Utilizing the introduced probability function, the expected cost of CPM is calculated as follows.

$$\mathbb{E}(CPM) = \sum_{J=0}^{\infty} \gamma(J)(\sum_{j=0}^{J-1}(CPM(j,J)(P(j) - P(j+1))) + CPM(J,J) * P(J)) \tag{11}$$

Here, $P(j+1)$ is the likelihood of making correct predictions about user location up to the (j+1)th cell crossing and is subtracted from the $P(j+1)$, since $CPM(j,J)$ calculates the cost when only the first j predictions are correct.

A closer look at (10) and (11) shows that if p is equal to 1 or 0 then the cost of the predictive mode would be minimized and equal to $\mathbb{E}[CPM(J,J)]$. Similarly, it can be shown that the cost of PrMM would be maximized when p is equal to 0.5. The only remaining issue is to find an upper bound for the J in (11) i.e. J_{max}. Considering a large random number for J_{max} makes the calculation of (11) susceptible to either underestimation of the upper bound when the ratio of the $\frac{\lambda_m}{\lambda_c}$ is high, or overestimation of that which in return causes unnecessary computation burden.

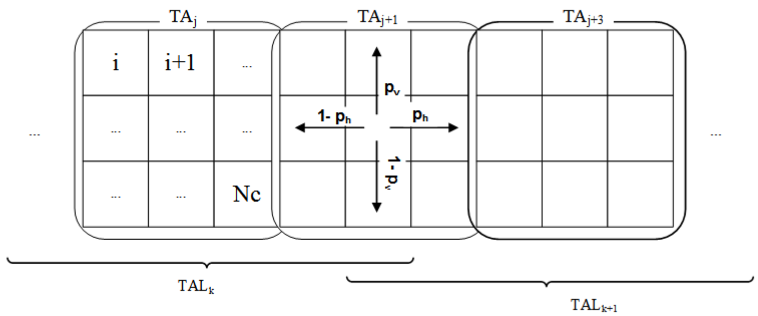


Fig. 6: configuration of the 2-D analytical framework

To resolve this problem, we first introduce a small error threshold term i.e. e for inaccuracies in estimation of the upper bound. Then J_{max} is estimated as follows.

$$J_{max} = \left\{ J | \int_0^J \gamma(j) > 1 - e \right\} \tag{12}$$

Substituting the value of the $\gamma(J)$ from (8) into (12) and rearranging the parameters, J_{max} is calculated as (13).

$$J_{max} = \left\lceil 1 - \frac{\log(a_0) * V * \lambda_m^2}{\log(V * \lambda_m * \lambda_c + 1)} \right\rceil \tag{13}$$

where,

$$a_0 = \frac{(e - 1 + a2) * \log(V * \lambda_m * \lambda_c + 1)}{a1 * V * \lambda_m} + 1 \tag{14}$$

$$a_1 = \left(\frac{\lambda_m}{\lambda_c}\right) \left(1 - (V * \lambda_c * \lambda_m + 1)^{(\frac{-1}{V * \lambda_m^2})}\right)^2 \tag{15}$$

$$a_2 = 1 - \left(\frac{\lambda_m}{\lambda_c}\right) \left(1 - (V * \lambda_c * \lambda_m + 1)^{(\frac{-1}{V * \lambda_m^2})}\right) \tag{16}$$

Through the next section, we evaluate the signaling cost of the predictive and conventional signaling schemes.

D. Extension to 2-D mobility model

Fig. 6 represents the configuration of the proposed framework when extended into 2D mesh cell configuration (i.e., Manhattan-street layout). Here UE is a 2-D random walk which makes horizontal movements with a probability of p_d and vertical movements with the probability of $(1 - p_d)$. In horizontal movements forward and backward moves are made with the probability of p_h and $1 - p_h$ and in vertical movements up and down displacements are with the probability of p_v and $1-p_v$. Given this configuration, the expected number of cell crossings before tracking area update i.e. N_{cr} (4) is modified as follows.

$$N_{cr}^{*} = \min\left\{\frac{N_{cr}}{p_d}, \frac{N_{cr}}{1-p_d}\right\} \tag{17}$$

The expected number of non-periodic updates i.e.N_{np} is also adapted as follows.

$$N_{np} = \frac{\lambda_m}{\lambda_c * N_{cr}^{*}} \tag{18}$$

The signaling cost of the conventional mode i.e. $\mathbb{E}(CM)$ will be similar to (1). The expected cost of the predictive model i.e. $\mathbb{CPM}$ is similar to (11) but p_m is calculated with considering both movement directions as follows.

$$p_m = \max\left\{p_d * p_h,\ p_d * (1-p_h),\ (1-p_d) * p_v,\ (1-p_d) * (1-p_h)\right\} \tag{19}$$

VIII. Numerical evaluation

This section investigates the performance of PrMM against the EMM signaling. In particular, we are interested in identifying those states of user mobility and configurations of network entities where the signaling cost of PrMM is less than EMM scheme. That is equivalent to identify the local maximums of the signaling reduction function R, which is defined as follows.

$$R = \left\{1 - \frac{\mathbb{E}(CPM)}{\mathbb{E}(CM)} | \lambda_m, \lambda_c, \alpha, N_c, N_t, p, C_s\right\} \tag{20}$$

Comparison of the maximum achievable signaling reduction in different situations gives a holistic view to the PrMM performance. Such analysis enables us to investigate the applicability of the PrMM in real world situations. Once the optimum points of the performance are calculated, a sensitivity analysis is conducted to evaluate the impact of the input parameters on signaling reduction value in the proximity of its optimum point. Section VIII-A explains how different parameters of the signaling reduction function are calculated. Section VIII-B presents the results and discussions. We start with the macro cells and continue to small cells. Since the observed trends of the signaling reduction in 1-D and 2-D evaluations were similar, we present a detailed analysis of the results from 1-D framework along with sample results from the 2-D evaluations.

A. Evaluation settings

To calculate the *signaling reduction* value, the cost functions of both standard MME signaling and the proposed mobility management are calculated according to (1) and (11). These functions

TABLE II: Simulation parameters

<table>
<tr><th colspan="2">Parameters</th><th colspan="2">Range/Value</th></tr>
<tr><td colspan="2">p</td><td colspan="2">0.55 (low predictability)
0.75 (medium predictability)
0.95 (high predictability)</td></tr>
<tr><td colspan="2">Idle period duration</td><td colspan="2">900-7200s</td></tr>
<tr><td colspan="2">Paging probability</td><td colspan="2">0.05-0.95</td></tr>
<tr><td colspan="2">TA size</td><td colspan="2">1-300</td></tr>
<tr><td colspan="2">TAL size</td><td colspan="2">1-16</td></tr>
<tr><td colspan="2">Cell size</td><td colspan="2">Small: r= 50 m
Macro: r=1000 m</td></tr>
<tr><td rowspan="4">CCR</td><td>CCR rate</td><td>Small cell</td><td>Macro cell</td></tr>
<tr><td>High</td><td>1800-2400</td><td>90-120</td></tr>
<tr><td>Medium</td><td>600-1800</td><td>30-90</td></tr>
<tr><td>Low</td><td>2-600</td><td>2-30</td></tr>
<tr><td rowspan="4">PA density</td><td>Level</td><td colspan="2">Number of cells</td></tr>
<tr><td>High</td><td colspan="2">3600-4800</td></tr>
<tr><td>Medium</td><td colspan="2">1200-3600</td></tr>
<tr><td>low</td><td colspan="2">1-1200</td></tr>
</table>

take network configuration parameters including N_c and N_t alongside with the user mobility characteristics including p, V, λ_m, paging probability α and mean idle state duration $\frac{1}{\lambda_c}$ as input and returns the signaling cost of the conventional and predictive signaling.

The LDC method is implemented to estimate the prediction update cost. The size of the required geohash code is set to 10 characters which can be represented with 50 bits. A geohash with 10 characters can represent the location information with the precision of approximately 1m, which suffices the positioning precision requirements for small cells. Table **??** summarizes the parameters that are used through the evaluation process and their allocated values. Herein, we categorize the independent parameters related to network configuration and mobility mode of the user into different ranges carefully selected to represent the typical real world situations.

For simplification λ_m and λ_c are combined and provided as expected Cell Crossing Rate CCR which is equivalent to $\frac{\lambda_c}{\lambda_m}$. The maximum and minimum allowable values for the expected CCR are calculated based on our assumption on the cell size, mobility speed of the user and duration of the idle period, each of which is specified in Table REFMISSING. Values between the maximum and minimum CCR are divided into high, medium and low rates. The maximum TAL size (i.e.

N_t) is set to 16 according to 3GPP specifications [29]. For TA size (i.e. N_c) we assumed that 300 cells would be a plausible cap; although 3GPP specifications allow for even a higher number of cells (e.g., Public Land Mobile Network (PLMN) can have up to 65,536 different cell-IDs) [6]. The combination of TAL and TA allows for generating paging area with up to 4800 cells. Table 1 divides the possible paging area sizes to three categories of high, medium and low densities. Through the evaluations the sizes of TA and TAL are determined based on the PA density level constraints. The range of paging probability (α) range and movement probability (p) are the same for all stimulation sets and are specified in Table 1.

The regions where the relative signaling cost of PrMM on EMM is minimized are identified for each unique parameters combination. For a given cell size, the evaluations are performed in three sets each concern a particular value of p. Each set comprises nine optimization processes that cover all possible combinations of CCR and PAs. Considering two cell size, three PA ranges, three CCR ranges and three p values, 54 maximum points are calculated in overall. We employed the Genetic algorithm [30] to locate the optimum points of the signaling reduction function from (20) over each given region of parameters.

B. Results and discussions

Figs. 7 and 8 represent the optimum points calculated for macro cells for different PA densities, CCR rates, and p values. It is evident that the signaling reduction degrades with the CCR rate.

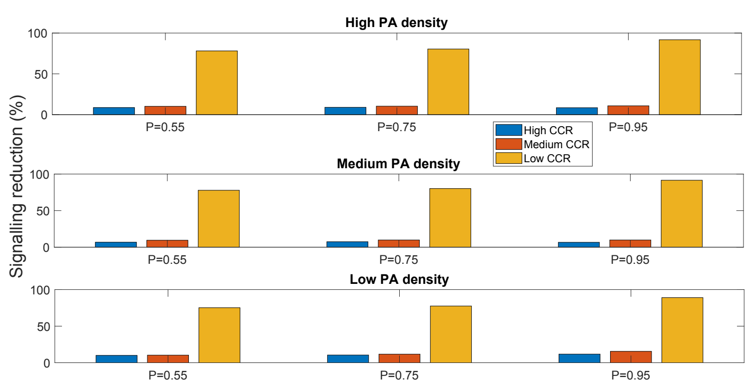


Fig. 7: Maximum signaling reductions in 1-dimension movement scenario calculated for macro cells, against F/B movement probability (p), PA density and CCR levels.

This could be down to the fact that at the higher CCR rates the PrMM updates accommodates more predictions and the overhead of the PrMM increases. Furthermore, longer sequences of predictions increase the likelihood of the prediction failure before the end of the idle period. This makes PrMM incur an extra signaling to update the location and switch to EMM scheme. The other noticeable observation in Fig.7 is that the higher values of the p improve the signaling reduction. This observation is aligned with our earlier discussion on equation (11), that was presented in Section VII-C. On the other hand, the PA density has a marginal effect on the signalling reduction of PrMM.

Fig. 8 represents the maximum signaling reduction values for small cells. Again, similar to macro cells, the signaling reduction is degraded with the CCR rate. Also, the results show that the signaling reduction improves with p values, which is analogous to the previous observations for macro cells. The optimum signaling reduction in small cells and macro cells are at a similar

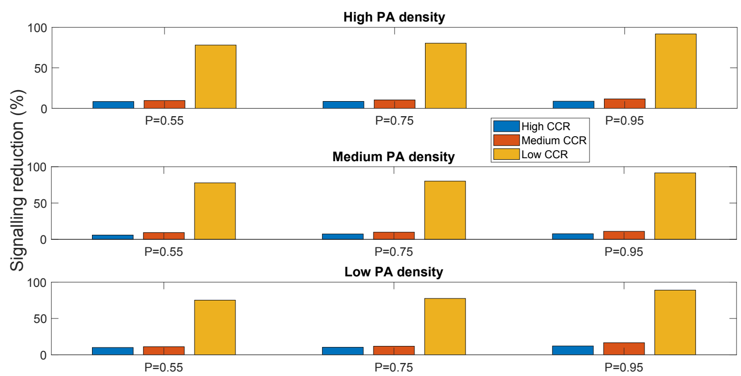


Fig. 8: Maximum signaling reductions in 1-dimension movement scenario calculated for small cells, against F/B movement probability (p), PA density and CCR levels.

level when CCR is low. However, this trend changes when CCR is set to medium and high rates. The discrepancy between results at different cell sizes is due to the fact that in the case of small cells the high or medium CCR causes a higher order of cell crossings with respect to the macro cells. This phenomenon is also evident in table.1. In this regard, PrMM has to generate a longer sequence of predictions which incur a higher load for the initial update. Also, a longer sequence of predictions increases the likelihood of prediction failure before the end of the idle mode.

Through the next set of evaluations, we study the impact of variation of the input parameters including N_c and N_t which represent the network configuration, and CCR and α that represent user behavior characteristics. A comparison between macro and small cells shows that similar trends and behaviors can be observed in both cases. For this reason, herein we confine our discussion to macro cells. Furthermore, for the purpose of demonstration, we use the results related to the p = 0.75 which is the middle range for p. The concluding remarks is applicable to other values of p.

Fig. 9 shows the results of sensitivity analysis over N_c and N_t values for each unique combination of CCR and PA levels. Here, at each row, the range of CCR increases from left to right and at each column, the range of PA density decreases from top to bottom. It is evident from

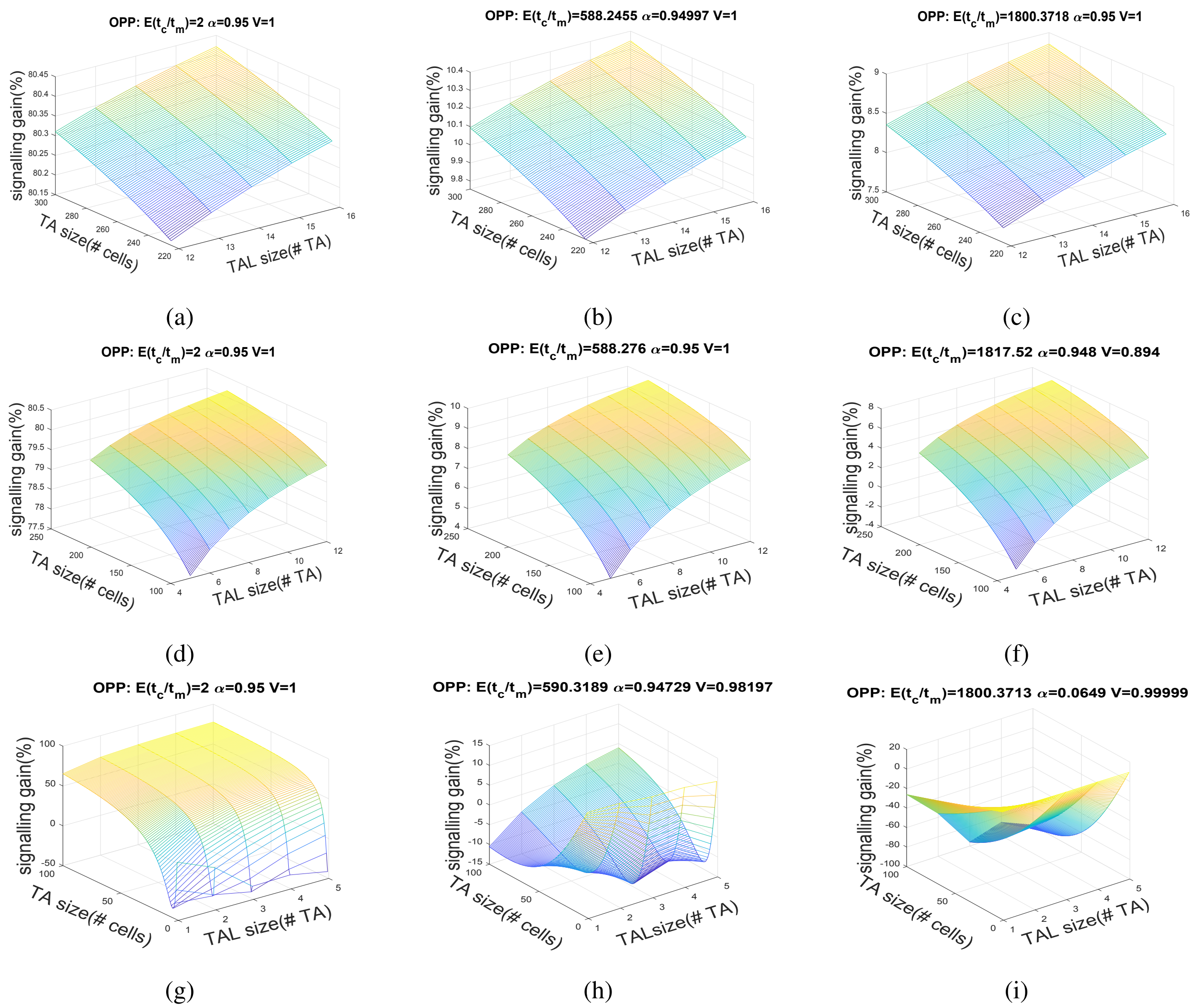


Fig. 9: Sensitivity analysis on PrMM performance on N_c and N_t for macro cells in 1-dimension movement scenario when p=0.75, the PA density level reduces from top to down and, CCR level reduces from left to right. Acronym OPP: Optimum Point Properties.

Fig. 9, that at each column, the sensitivity to the N_c and N_t values increases with reducing the PA density. For example, in the first column, at fig. 9.a which represents the lowest PA density the variation of signaling reduction on N_c and N_t values is less than 1%, while in Fig.9.g the signaling reduction variation reaches 40%. The similar trend can be observed in other columns.

One explanation for the stability of the signaling reduction at the first row could be the fact that the overhead of TAU, which is dependent on the sizes of tracking area and tracking area list, is negligible in comparison with paging. By reduction of PA density in the next rows, the impact of TAU becomes notable.

The variation of CCR from left to right shows that in the vicinity of the optimum points the signaling reduction improves at lower CCRs. Comparing the result of CCR on the sensitivity of the signaling reduction to the N_t and N_c parameters can be studied by comparing the columns. It can be seen that at lower CCR rates (i.e. $\mathbb{E}(t_c/t_m)$) the variation of signaling reduction with respect to the N_c and N_t parameters increases. This observation is more clear when PA density is at its lowest level. Again the reason is related to the TAU overhead which is highly dependent on the cell crossing rate and its overhead impact is more evident at lower PA densities. Results in Fig. 9 implies that high PA density and lower CCR provides the most appropriate condition for switching into PrMM.

Fig. 10, concerns the variation of the signaling reduction with respect to different values of CCR and paging probability. Reviewing columns of fig. 10 from left to right shows that the signaling reduction monolithically improves by reduction of the CCR. Also, signaling reduction sensitivity on paging probability reduces at lower ranges of CCR. These observations confirm our earlier deductions that when the number of predictions is high (i.e. higher CCR rates), the impact of PrMM would be mainly related to reduction of the paging costs; and moving towards lower CCR ranges put more weights on the impact of the TAU overhead reduction. A row-wise analysis of Fig. 10 reveals a remarkable stability in signaling reduction behavior on paging probability and CCR at different PA densities. The signaling reduction also improves with paging probability however the variations become less noticeable when paging probability is above 15%.

C. 2-D evaluation sample results

This section evaluates performance of PrMM in 2-dimensions mobility scenarios. Figs. 11 and 12 shows the signalling reduction gain in macro cell and small cell networks, respectively, while the sensitivity analysis results are provided in Figs. 13 and 14.

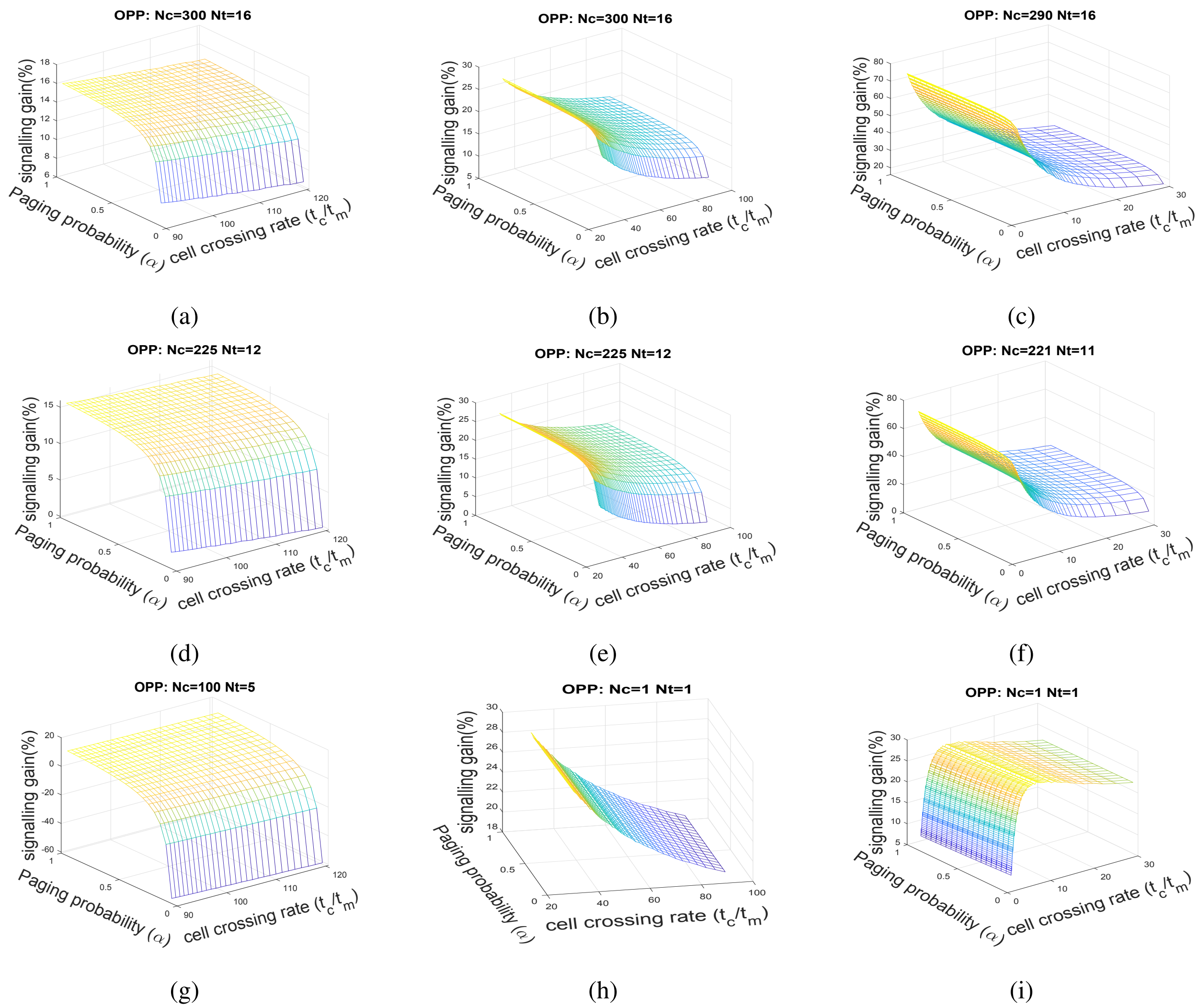


Fig. 10: Sensitivity analysis on PrMM performance on CCR and α for macro cells in 1-dimension movement scenario when $p = 0.75$, the PA density level reduces from top to down and, CCR level reduces from left to right. Acronym OPP: Optimum Point Properties.

It can be observed that the PrMM signalling reduction gains in the 2-dimensions movement scenario (Figs. 11 and 12) are significantly less than the gains of the 1-dimension movement scenario (Fig. 7 and 8). This can be traced to the fact that the prediction accuracy is significantly less in the former case, i.e., there are more possible movement directions in the 2-dimensions case. As with the 1-dimension movement case, the signalling reduction in the 2-dimensions movement case is inversely proportional to the CCR. As shown in Figs. 11 and 12, the PrMM increases the signalling load (i.e., the signalling reduction becomes negative) in high and medium

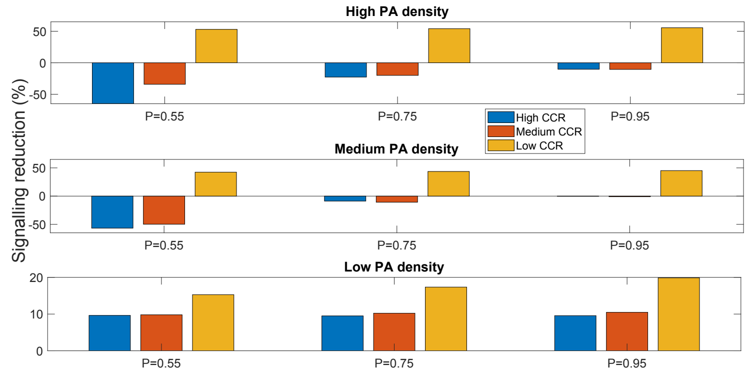


Fig. 11: Maximum signaling reductions in 2-dimensions movement scenario calculated for macro cells, against F/B movement probability (p), PA density and CCR levels.

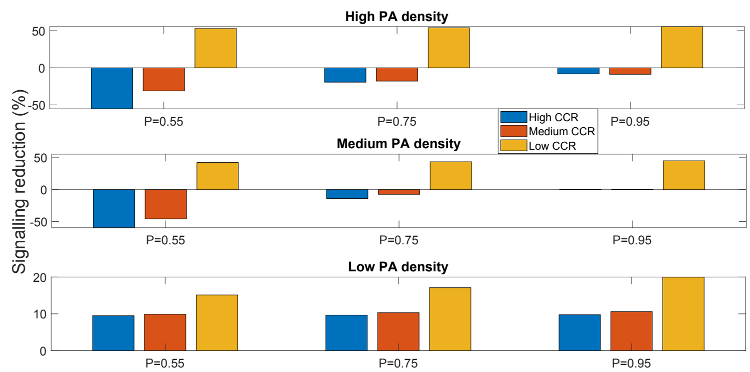
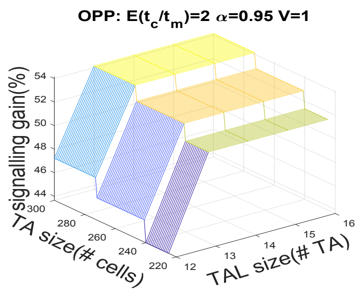


Fig. 12: Maximum signaling reductions in 2-dimensions movement scenario calculated for small cells, against F/B movement probability (p), PA density and CCR levels.

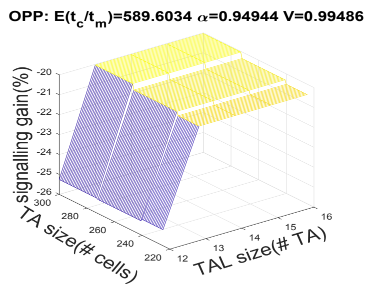
CCR, with high and medium PA density. This can be linked to the higher probability of error in the 2-dimensions movement case coupled with the longer sequence of prediction and the higher overhead and signalling in the high/medium CCR scenario. In low PA density, the PrMM reduces the signalling load (i.e., the signalling reduction is positive), however, these gains are significantly less than the peak gains in the high PA density with low CCR case. Expressed differently, the switching towards PrMM provides the peak gains in both the 1-dimension and the 2-dimension movement scenarios when the PA density is high and the CCR is low.

IX. Conclusion

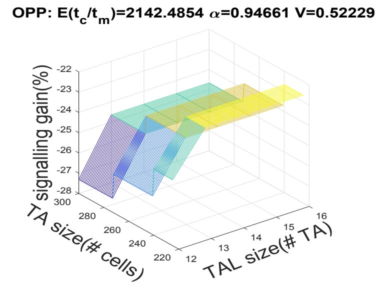
The increasing overhead of mobility management signaling has become a major challenge for the next generation of mobile communication networks. Predictive Mobility Management is



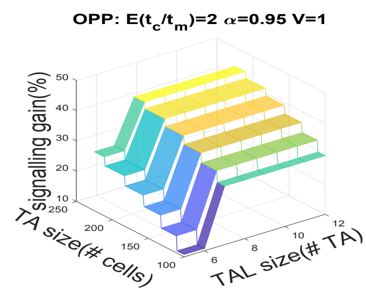
(a)



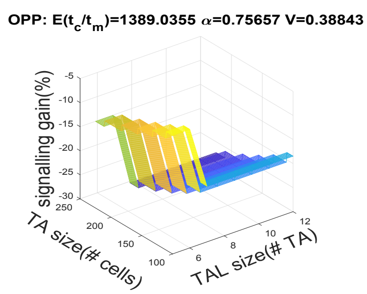
(b)



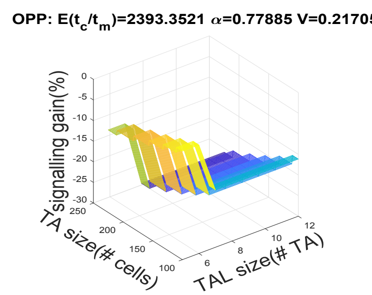
(c)



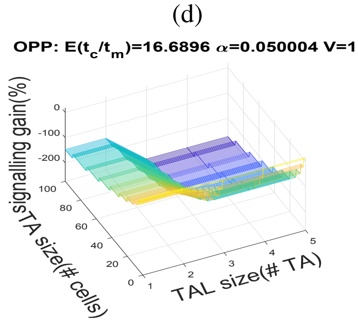
(d)



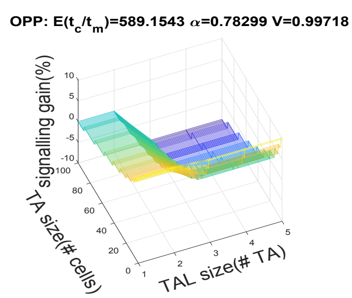
(e)



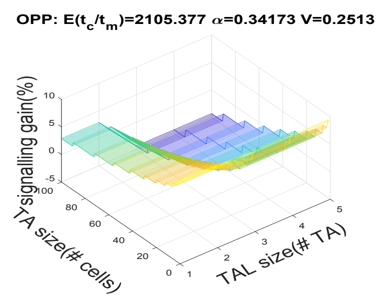
(f)



(g)



(h)



(i)

Fig. 13: Sensitivity analysis on PrMM performance on N_c and N_t for macro cells in 2-dimension movement scenario when p=0.75, the PA density level reduces from top to down and, CCR level reduces from left to right. Acronym OPP: Optimum Point Properties.

a new trend for more efficient mobility management in next generation mobile communication networks. However, existing predictive solutions either deteriorate critical KPIs such as the connection establishment time or add a significant computation load to the network entities. Moreover, the applicability of such solutions on different user mobility characteristics and network configuration is not sufficently studied. In this work, we propose PrMM as a new architecture for predictive mobility management which addresses some of the major shortcomings of the existing solutions. To study the characteristics of the proposed solution, we developed an

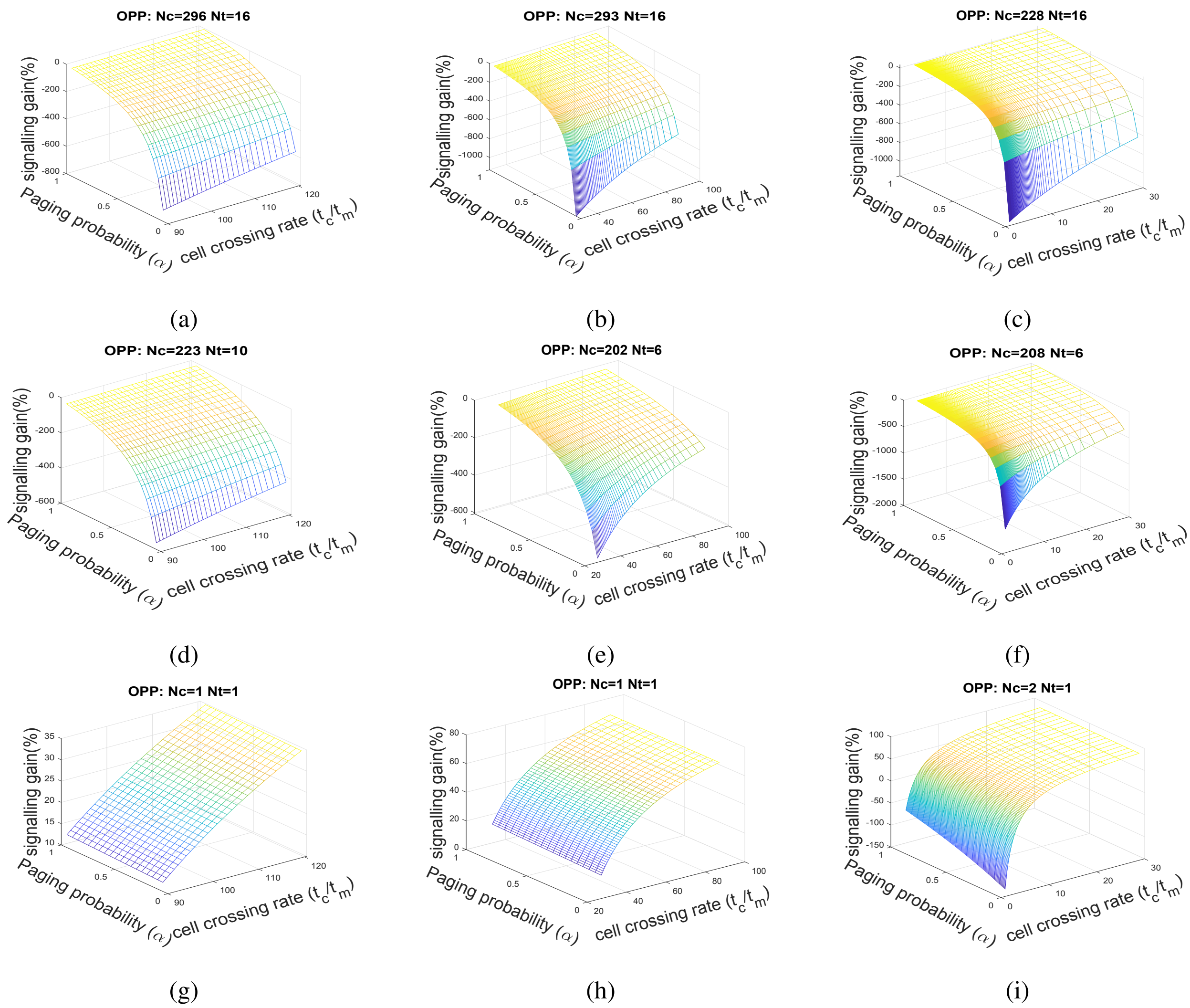


Fig. 14: Sensitivity analysis on PrMM performance on CCR and α for macro cells in 2-dimensions movement scenario when $p = 0.75$, the PA density level reduces from top to down and, CCR level reduces from left to right. Acronym OPP: Optimum Point Properties.

analytical framework which incorporates both user mobility characteristics and network entities arrangements to calculate the expected signaling overhead of both PrMM and standard EMM scheme.

Relative overhaed of PrMM on EMM is evaluated through several numerical tests. The evaluation results confirm our intuition that the cost of PrMM is inversely proportional to the predictability of the UE mobility. In particular, our results suggest that in the vicinity of the optimum points, the signaling reduction is a monotonic function of the cell crossing rate and paging probability. Signaling reduction is also a monotonic function of PA density at the higher ranges of PA densities and becomes a concave function of that at the lower ranges.

Depending on user mobility and network configuration, the PrMM efficiency has been shown to surpass the conventional signaling scheme by over 90%. The analysis of the correlation between signaling reduction and different aspects of user mobility and network configuration indicates that the best performance is typically achieved when the density of the paging area is at its highest level and cell crossing rate is relatively low. Also, in such situation, the signaling reduction is almost independent of the paging probability. This observation is particularly important considering the fact that ultra-dense small cells are expected to play a significant role in the next generation of communication networks. Our analysis shows that if an appropriate arrangement of N_c and N_t is provided PrMM is capable of improving on signaling overhead in almost any mobility mode of the user and paging probability. This observation could be a starting point for methods emerging from network function virtualization that consider to adapt network parameters to user behavior or in self-organizing networks. PrMM has been sensitive to unreliable predictions at high CCR rates, particularly when the PA density is low. In this regard, PrMM agents at UE and MME should consider the network configuration and prediction confidence level before switching to predictive mode. Realization of this feature remains for our future works.

References

[1] G. Nowoswiat, "Managing lte core network signaling traffic," 2014. [Online]. Available: http://www2.alcatel-lucent.com/techzine/managing-lte-core-network-signaling-traffic/

[2] ——, "Managing lte core network signaling traffic," 2014. [Online]. Available: https://insight.nokia.com/manage-signaling-impact-small-cells

[3] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010. [Online]. Available: http://science.sciencemag.org/content/327/5968/1018

[4] S.-M. Qin, H. Verkasalo, M. Mohtaschemi, T. Hartonen, and M. Alava, "Patterns, entropy, and predictability of human mobility and life," *PLOS ONE*, vol. 7, no. 12, pp. 1–8, 12 2012. [Online]. Available: https://doi.org/10.1371/journal.pone.0051353

[5] 3GPP, "Evolved universal terrestrial radio access (e-utra) and evolved universal terrestrial radio access network (e-utran)," *Technical Specification 3GPP TS 36.300*, vol. 11.10.0, 2014.

[6] ——, "Non-access-stratum (nas) protocol for evolved packet system (eps)," *Tech. Spec. 3GPP TS 24.301*, vol. 9.3.0, 2010.

[7] R. H. Liou, Y. B. Lin, and S. C. Tsai, "An investigation on lte mobility management," *IEEE Transactions on Mobile Computing*, vol. 12, no. 1, pp. 166–176, Jan 2013.

[8] 3GPP, "Paging load reduction in asynchronous cells," *Tech. Rep. 3GPP R3-081412 of TSG-RAN-WG3 Meeting 60*, 2007.

[9] A. Roy, J. Shin, and N. Saxena, "Entropy-based location management in long-term evolution cellular systems," *IET Communications*, vol. 6, no. 2, pp. 138–146, January 2012.

[10] B. Liang and Z. J. Haas, “Predictive distance-based mobility management for multidimensional pcs networks,” *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, pp. 718–732, Oct 2003.

[11] A. Mohamed, O. Onireti, M. A. Imran, A. Imran, and R. Tafazolli, “Predictive and core-network efficient rrc signalling for active state handover in rans with control/data separation,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1423–1436, March 2017.

[12] H. Wang, S. Chen, M. Ai, and H. Xu, “Localized mobility management for 5g ultra dense network,” *IEEE Transactions on Vehicular Technology*, vol. 66, no. 9, pp. 8535–8552, Sept 2017.

[13] S. M. Razavi, D. Yuan, F. Gunnarsson, and J. Moe, “Dynamic tracking area list configuration and performance evaluation in lte,” in *2010 IEEE Globecom Workshops*, Dec 2010, pp. 49–53.

[14] ——, “Performance and cost trade-off in tracking area reconfiguration: A pareto-optimization approach,” *Computer Networks*, vol. 56, no. 1, pp. 157 – 168, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128611003331

[15] S. Jeon, S. Figueiredo, R. L. Aguiar, and H. Choo, “Distributed mobility management for the future mobile networks: A comprehensive analysis of key design options,” *IEEE Access*, vol. 5, pp. 11 423–11 436, 2017.

[16] H. Ko, G. Lee, S. Pack, and K. Kweon, “Timer-based bloom filter aggregation for reducing signaling overhead in distributed mobility management,” *IEEE Transactions on Mobile Computing*, vol. 15, no. 2, pp. 516–529, Feb 2016.

[17] J. H. Winters, “Device location prediction for mobile service optimization,” April 2014, uS Patent 8,706,131.

[18] J. S. Csapo, R. J. Webb, and P. Y. Sohn, “Apparatus and method for mobile station-assisted optimization of a wireless network,” May 2012, uS Patent 8,190,145.

[19] B. Hogan, “Use of mobile locating and power control for radio network optimization,” August 2002, uS Patent 6,442,393.

[20] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, “Big data-driven optimization for mobile networks toward 5g,” *IEEE network*, vol. 30, no. 1, pp. 44–51, 2016.

[21] S. A. Hoseinitabatabaei, A. Gluhak, R. Tafazolli, and W. Headley, “Design, realization, and evaluation of udirect-an approach for pervasive observation of user facing direction on mobile phones,” *IEEE Transactions on Mobile Computing*, vol. 13, no. 9, pp. 1981–1994, Sept 2014.

[22] A. Mohamed, M. A. Imran, P. Xiao, and R. Tafazolli, “Memory-full context-aware predictive mobility management in dual connectivity 5G networks,” *IEEE Access*, vol. 6, pp. 9655–9666, 2018.

[23] K.-H. Chiang and N. Shenoy, “A 2-d random-walk mobility model for location-management studies in wireless networks,” *IEEE Transactions on Vehicular Technology*, vol. 53, no. 2, pp. 413–424, March 2004.

[24] L. Sanguinetti *et al.*, “Large system analysis of the energy consumption distribution in multi-user mimo systems with mobility,” *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1730–1745, March 2015.

[25] A. Orsino *et al.*, “Time-dependent energy and resource management in mobility-aware d2d-empowered 5g systems,” *IEEE Wireless Communications*, vol. 24, no. 4, pp. 14–22, 2017.

[26] 3GPP, “Lte radio access network (ran) enhancements for diverse data applications,” *Tech. Rep. TR 36.822*, vol. 11, 2012.

[27] T. Sandholm, “Geographically partitioned online content services,” Mar. 21 2013, uS Patent App. 13/233,878. [Online]. Available: https://www.google.ch/patents/US20130072233

[28] A. Roy, S. K. Das, and A. Misra, “Exploiting information theory for adaptive mobility and resource management in future cellular networks,” *IEEE Wireless Communications*, vol. 11, no. 4, pp. 59–65, Aug 2004.

[29] 3GPP, “E-UTRAN; S1 Application Protocol (S1AP),” *Tech. Spec. TS 36.413*, vol. 12.3.0, 2014.

[30] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989.