Mobility Prediction based Autonomous Proactive Energy Saving (AURORA) Framework for Emerging Ultra-Dense Networks

Hasan Farooq, Ahmad Asghar and Ali Imran, *Member, IEEE,*

***Abstract*—Increased network wide energy consumption is a paramount challenge that hinders wide scale ultra-dense networks (UDN) deployments. While several Energy Saving (ES) enhancement schemes have been proposed recently, these schemes have one common tenancy. They operate in reactive mode i.e., to increase ES, cells are switched ON/OFF reactively in response to changing cell loads. Though, significant ES gains have been reported for such ON/OFF schemes, the inherent reactiveness of these ES schemes limits their ability to meet the extremely low latency and high QoS expected from future cellular networks vis-a-vis 5G and beyond. To address this challenge, in this paper we propose a novel user mobility prediction based AUtonomous pROactive eneRgy sAving (AURORA) framework for future UDN. Instead of observing changes in cell loads passively and then reacting to them, AURORA uses past hand over (HO) traces to determine future cell loads. This prediction is then used to proactively schedule small cell sleep cycles. AURORA also incorporates the effect of Cell Individual Offsets (CIOs) for balancing load among cells to ensure QoS while maximizing ES. Extensive system level simulations leveraging realistic SLAW model based mobility traces show that AURORA can achieve significant energy reduction gain without noticeable impact on QoS.**

***Index Terms*—5G, Energy Saving, Mobility Prediction, Proactive SON, Heterogeneous Networks, Sleeping Cells, ON/OFF Small Cells, CIOs.**

I. Introduction

The current exponential mobile data traffic escalation is a precursor towards an imminent "capacity crunch". In this backdrop, extreme network densification through deployment of large number of Small Cells (SCs) has emerged as the most yielding solution to achieve the 1000 fold capacity gain goal [1]. However, the ultra-dense deployments of SCs is on direct collision path with the economically viable and energy efficient deployment vision of 5G. This is due to the high aggregated network energy that "always ON" small cells are bound to consume in an Ultra Dense Network (UDN). In addition to higher carbon footprint, this translates into higher OPEX. Although SCs have a relatively lower power consumption profile, yet the always ON approach increases overall network wide energy consumption [2]. This is because the load independent power consumption (circuit power) component in SCs constitutes a much larger portion of over-all power consumption [3]. As a result, with advent of UDN, the need for ES schemes will be even more compelling. It is a consensus among research community that to avert possible energy crunch in 5G and to achieve economic viability, the $1000\times$ capacity increase must be achieved at a similar or lower power consumption as legacy networks [4].

H. Farooq, A. Asghar and A. Imran are with the Department of Electrical and Computer Engineering, University of Oklahoma, Tulsa, OK, 74135 USA e-mail: (hasan.farooq@ou.edu, ahmad.asghar@ou.edu, ali.imran@ou.edu, see http://www.ai4networks.com for complete information). The core idea of AURORA framework has won IEEE GREEN ICT Best Solution Award Competition 2017.

A. Related Work

Energy consumption in cellular systems can be reduced significantly by turning OFF underutilized cells during off-peak hours or by optimizing resource allocation such that minimum energy is consumed per bit transmission [4]–[7]. To exploit these approaches recently ES has been adopted as a key Self Organizing Network (SON) function by 3GPP [8] and has been extensively studied in literature. ES enhancement with focus on optimizing resource allocation despite of its relatively small gain compared to turning ON/OFF under-utilized BSs has been studied more extensively compared to later approach [4]. The resource allocation optimization can reduce the energy consumption to only a limited degree for a given system throughput target. ES of the cellular systems can be further enhanced significantly by switching under-utilized BSs to sleep mode or turning them OFF entirely during off-peak time [5]–[7], [9]. In this direction of research, some recent works show promising results in terms of potential ES [10]–[23]. However, to the best of our knowledge, existing ES approaches fall short of mark for 5G requirement due to following four limitations:

1) **Reactive mode of operation**: Conventional ES SON algorithms are designed to switch OFF/ON cells after detecting network conditions that have already taken effect. For example, when congestion is detected in network, usually a non-convex NP-hard ES algorithm is solved to identify certain sleeping/OFF cells, that should be switched ON or using same process certain cells are switched OFF, when low load is observed in certain cells. This is an improvement over fixed timer based switching ON/OFF [24] that can at best follow a coarse statistical spatio-temporal traffic pattern and thus achieves ES at cost of QoS. However, given the acute dynamics of traffic and cellular environment, by the time congestion or low traffic conditions are detected and a realistic non-convex NP-hard ES algorithm is solved to produce new network ON/OFF configuration optimal for observed network conditions, the conditions may already change. Thus, the newly determined switch ON/OFF vector is likely to be suboptimal before it can be actuated. This problem can exacerbate particularly in 5G, where a

motely of traffic and plethora of cell types means the dynamics of cellular eco-system will be even more swift.

2) **Difficulty in meeting 5G low latency:** Base Stations require a certain amount of time to wake up from sleep cycle [25]. For a user entering a sleeping cell, this time to wake up will add to the latency experienced by the user. This demands paradigm shift from the conventional reactive design of ES algorithms towards proactive characteristics to cope with extreme low latency requirements of 5G in a more agile fashion.
3) **Impractical cell discovery:** A key challenge in switching OFF based ES schemes is: how to discover an OFF cell when users enter into physical coverage area of the OFF cell? Existing ES schemes either overlook this challenge, or propose solutions that either exploit neighboring cells or a master controller to wake up the cell, when enough users enter into the coverage area of OFF cell. This approach may work in low user density network with large macro cells with relatively less stringent Quality of Service (QoS) requirements such as LTE, but it may not scale to 5G because of signaling overhead, delays and cost of missing out OFF small cells for off-loading.
4) **SON Conflict prone design:** The other caveat with conventional ES solutions is that they are oblivious of the fact that multiple SON functions may be prone to hidden or undesired conflict when implemented together in a network [1], [26]. Two SON use cases that become highly relevant to the ES in HetNets are Coverage and Capacity Optimization (CCO) and Load Balancing (LB) [8] because of the overlap among their optimization parameter set: Transmission Power and Cell Individual Offsets (CIOs). When an ES switches OFF some cells, it may force some users to be associated to neighboring ON cells and overload them thereby conflicting with CCO and LB SON functions. As explicated in [26], such conflict prone ES solution design can actually degrade network's performance instead of improving it.

B. Contributions and Organization

To address the aforementioned limitations, we propose AURORA framework (Fig. 1) by building on the lines of Big Data empowered SON framework [1]. The key idea is to make emerging cellular systems artificially intelligent and autonomous so that they can anticipate user mobility behavior. This intelligence in turn is then used to formulate a novel ES optimization problem that proactively schedules small cell sleep cycles to divert and focus the right amount of resources when and where needed while satisfying QoS requirements. The contributions and organization of paper can be summarized as follows:

1) As a building block of AURORA, we develop and analyze a Semi-Markov model based spatio-temporal mobility prediction framework. Our proposed mobility prediction model overcomes the limitation of conventional discrete time Markov chain based prediction models that fail to incorporate time dimension i.e., "Time of next HO" (Section II-B). Next, we propose a novel method to map the next cell spatiotemporal HO information to the estimated future location coordinates based on the idea of Landmarks (Section II-C). This novel method further increases the spatial resolution of the future location estimation without requiring increase in number of states for Semi-Markov model. The accuracy of proposed model is quantified through extensive Monte Carlo simulations.
2) Based on the intelligence gained from the mobility model i.e., future cell loads, a proactive energy saving optimization problem is formulated to minimize the energy consumption by switching OFF underutilized SCs (Section II-D). In addition to proactivness, another key novelty of proposed ES scheme is that it leverages CIOs as optimization variables for balancing load among cells while deciding which cells to switch ON/OFF. In this way, an additional UDN specific mechanism is exploited to ensure QoS while maximizing ES. Although the formulated problem is non-convex large scale combinatorial and NP-hard, our results show that the structure of the problems allows heuristics such as genetic programming to find good solutions with high ES yield. The ahead of time estimation of cell loads allows ample time for such heuristics to converge without jeopardizing QoS.
3) We conduct multi-tier system level 3GPP compliant rigorous simulations for comprehensive performance analysis of proposed AURORA (Section III). The prediction accuracy of the Semi-Markov based mobility prediction model has been quantified using realistic SLAW mobility model in HetNets environment. The average location estimation error was found to be around 28 meters on average, while relying only on one piece of information that is already available in network i.e., HO trace.
4) We also analyze the impact of cell load thresholds on ES gains and QoS (percentage of satisfied users) for proactive energy saving optimization. The results of this analysis provide actionable insights for determining cell load thresholds that can judiciously strike the intended balance among the conflicting goals of ES and QoS.
5) We perform a comparative analysis of proposed solution, in Low and High Traffic demand scenarios with the latter comprising of all video users, against several bench marks including industrial practices i.e., All ON SCs without and with fixed CIOs. AURORA achieved 68% and 99% gain in the total network energy reduction for low and high traffic demand scenarios respectively by putting under-utilized SCs in sleep mode with negligible number of unsatisfied users. Moreover, we compare AURORA with near-optimal performance bound that is achievable when future network load conditions can be estimated with 100% accuracy. This comparison demonstrates that AURORA is reasonably resilient to location estimation inaccuracies.

II. AURORA Framework

In this section we present the analytical model development of AURORA Framework whose three key corner stones are:

- Semi-Markov Process based Spatiotemporal Next Cell Prediction

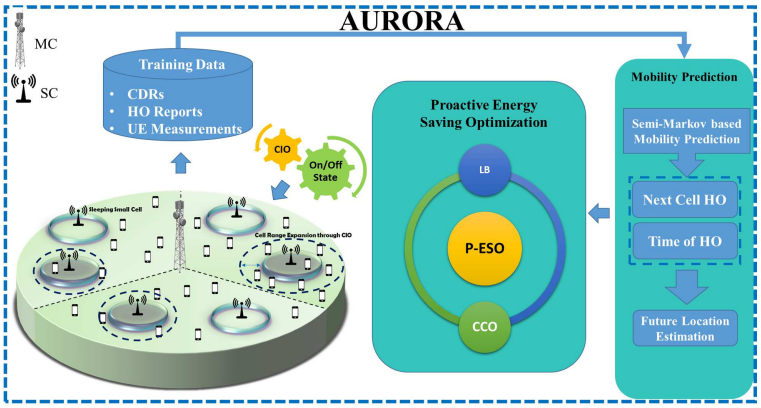


Fig. 1. AURORA Framework

- Mapping of Next Cell Prediction to Future User Location Estimation
- Proactive-Energy Saving Optimization based on Future User Location Estimation

A. Network Model and Assumptions:

The AURORA framework proposed in this paper only focuses on the downlink of cellular systems for the sake of conciseness. It is assumed that all mobile devices and small cells have omnidirectional antennas with a constant gain in all directions while macro cells have directional antennas. Frequency reuse of one is considered and same band is utilized by the macrocell and the small cells. A full buffer traffic model is used for each user, i.e., there is always data available to be sent for a user with constant bit rate service. A centralized C-SON architecture is assumed wherein a centralized server in the core network performs system wide Proactive-Energy Saving Optimization. Moreover, HO traces that include location stamped information of past cell transitions such as cell IDs, RSRPs and call detail records are assumed to be available to the C-SON server.

B. Semi-Markov based Spatiotemporal Next Cell Prediction

1) Background: Our rationale to build and utilize mobility prediction as a foundation for AURORA is backed by landmark study that analyzed real data for 10 million mobile users [27] and showed that typical human mobility features 93% average predictability. The mobility prediction model developed in this work builds on our recent study validated in real network [28] that exploits following idea: transition probability to a next cell can be predicted by modelling user transition from one cell to another as a Markov stochastic process and using HO history to estimate state transition probabilities. Discrete Time Markov Chain (DTMC) has been commonly used in the literature for mobility prediction purposes [29]–[31]. As compared to more complex and more space-consuming compression based predictors, the Markov based scheme can yield more scalable solution as it does not need to store users' past movements. Instead the crux of this information is captured by transition probabilities. However, DTMC is memory less and assumes sojourn time is geometrically distributed and each transition takes place in one unit time. Considering these limitations of the DTMC model, the aforementioned works have utilized DTMC for only the spatial prediction i.e., identification of future cell only without any information about the time at which handover may take place. Continuous Time Markov Chain (CTMC) is continuous counter part of DTMC and can be utilized for mobility prediction if the human mobility is assumed to be memory less and cell sojourn time is assumed to be exponentially distributed. As per [32], human mobility exhibits memory property and can be best approximated with power law (heavy tailed) distribution instead of memory less exponential distributions. Fortunately, Semi-Markov is an advanced class of Markov models that allows for arbitrary distributed sojourn times. Few recent works have characterized prediction accuracy performance of Semi-Markov based model for mobility prediction [33], [34]. However, to the best of our knowledge, this study is the first of its kind that presents spatio-temporal mobility prediction model, and a framework to transform that prediction into future cell load estimates. It then uses those load estimates to devise and analyze a proactive and QoS aware energy saving solution.

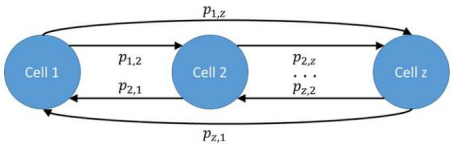


Fig. 2. Probability State Transition Diagram

2) Mobility Prediction Model: We begin by modeling user mobility as a Semi-Markov renewal process $\{(X_n, T_n) : n \geq 0\}$ with discrete state space $\mathbb{C} = 1, 2, 3 \ldots, z$ where T_n is the time of nth transition, X_n is the state at nth transition and total of z cells [28]. Each cell is represented by the state of the Semi-Markov process, and a handover from one cell to another is considered as state transition. It is assumed that the process is time-homogeneous during the time period in which the model is built. Fig. 2 shows state transition diagram for the Semi-Markov model wherein $p_{i,j}$ is the probability of transition from cell i to j. The associated time-homogeneous Semi-Markov kernel for user 'u' which is the probability of transition to j^{th} cell if user has already spent time t in i^{th} cell is defined as:

$$\psi_{i,j}^{(u)}(t) = Pr(X_{n+1}^{(u)} = j, T_{n+1}^{(u)} - T_n^{(u)} \leq t | X_n^{(u)} = i) \quad (1)$$

$$= p_{i,j}^{(u)} S_{i,j}^{(u)}(t) \quad (2)$$

where

$$p_{i,j}^{(u)} = \lim_{t\to\infty} \psi_{i,j}^{(u)}(t) \tag{3}$$

$$= Pr(X_{n+1}^{(u)} = j | X_n^{(u)} = i),\ p_{i,j}^{(u)} \in P^{(u)} \tag{4}$$

and

$$S_{i,j}^{(u)}(t) = Pr(T_{n+1}^{(u)} - T_n^{(u)} \leq t | X_{n+1}^{(u)} = j, X_n^{(u)} = i) \tag{5}$$

Here $p_{i,j}^{(u)}$ is the probability of handover of user 'u'from cell i to j, $\mathbf{P}^{(u)}$ is the probability transition matrix of the embedded Markov chain of user 'u'given as

$$\mathbf{P}^{(u)} = \begin{bmatrix} p_{1,1}^{(u)} & p_{1,2}^{(u)} & \cdots & p_{1,z}^{(u)} \\ p_{2,1}^{(u)} & p_{2,2}^{(u)} & \cdots & p_{2,z}^{(u)} \\ \vdots & \vdots & \vdots & \vdots \\ p_{z,1}^{(u)} & p_{z,2}^{(u)} & \cdots & p_{z,z}^{(u)} \end{bmatrix} \tag{6}$$

and $S_{i,j}^{(u)}(t)$ is the sojourn time distribution of user 'u'in cell i when next cell is j. It is important to note here that handover from cell to itself is not allowed, therefore diagonal of the matrix $P^{(u)}$ will be all zeros and the matrix will be a hollow matrix. Furthermore, direct handovers are possible between neighboring cells only. The probability that the user 'u' in cell i will leave cell i before or at time t regardless of the next cell is defined as:

$$\Lambda_i^{(u)}(t) = Pr(T_{n+1}^{(u)} - T_n^{(u)} \leq t | X_n^{(u)} = i) \tag{7}$$

$$= \sum_{j=1}^{z} \psi_{i,j}^{(u)}(t) \tag{8}$$

Now the time-homogeneous Semi-Markov process of user 'u' is defined as $X = (X_t,\ t \in \mathbf{R_0^+})$ with state transients as:

$$\phi_{i,j}^{(u)}(t) = Pr(X_t^{(u)} = j | X_0^{(u)} = i) \tag{9}$$

$$= (1 - \Lambda_i^{(u)}(t))\delta_{i,j} + \sum_{m=1}^{z} \int_0^t \phi_{m,j}^{(u)}(t-\tau) d\psi_{i,m}^{(u)}(\tau) \tag{10}$$

$$= (1 - \Lambda_i^{(u)}(t))\delta_{i,j} + \sum_{m=1}^{z} \int_0^t \frac{d\psi_{i,m}^{(u)}(\tau)}{d\tau} \phi_{m,j}^{(u)}(t-\tau) d\tau \tag{11}$$

where $\delta_{i,j}$is the Kronecker function defined as:

$$\delta_{i,j} = \begin{cases} 0 & , i \neq j \\ 1 & , i = j \end{cases} \tag{12}$$

Integral equations (10) and (11) are Volterra equations of the first and second kind and the integral is the convolution of $\psi_{i,m}^{(u)}(.)$ and $\phi_{m,j}^{(u)}(.)$ i.e., $\psi_{i,m}^{(u)} * \phi_{m,j}^{(u)}$. It gives the probability that user 'u' starting in cell i will be in cell j by t. The first part of the right-hand side is the probability that the user, being in cell i, never leaves cell i until the end of the period t. The second part of the right-hand side of equation accounts for all cases in which the transition from i to j occurs via another cell $m \neq i$ applying the renewal argument. First, the probability that the user stays in cell i for a period of length τ and then goes to cell m is given by $\psi_{i,m}^{(u)}(\tau)$. Handover to this new cell m can be interpreted as a renewal of the process because the expected behavior of the user from then on is the same irrespective of when the user enters cell m. Therefore, the probability that the user which is in cell m at τ will be in cell j at t is given by $\phi_{m,j}^{(u)}(t-\tau)$. As the transition from i to m can occur anytime between 0 and t, therefore all possible transition times are considered by the integration over τ [35]. The numerical solution to solve evolution equations (10) and (11) is given by [36] and we implement the same approach. The evolution equation (10) can be re-written for discrete-time homogeneous Semi-Markov process as:

$$\phi_{i,j}^{(u)}(k) = h_{i,j}^{(u)}(k) + \sum_{m=1}^{z} \sum_{\tau=1}^{k} \sigma_{i,m}^{(u)}(\tau) \phi_{m,j}^{(u)}(k-\tau) \tag{13}$$

where $h_{i,j}^{(u)}(k) = (1 - \Lambda_i^{(u)}(t))\delta_{i,j}$ and $\sigma_{i,m}^{(u)}(k) = \frac{d\psi_{i,m}^{(u)}(\tau)}{d\tau}$ can be approximated as follows assuming time step is equal to the unit:

$$\sigma_{i,m}^{(u)}(k) = \begin{cases} \psi_{i,m}^{(u)}(1) & , k = 1 \\ \psi_{i,m}^{(u)}(k) - \psi_{i,m}^{(u)}(k-1) & , k > 1 \end{cases} \tag{14}$$

As $\mathbf{P}^{(u)}$ is right stochastic matrix therefore $\psi^{(u)}(k)$ and $\phi^{(u)}(k)$ will also be a right stochastic matrices i.e., $\sum_{j=1}^{z} \psi_{i,j}^{(u)}(k) = \sum_{j=1}^{z} \phi_{i,j}^{(u)}(k) = 1, \forall i,j \in \mathbb{C}$. The $\phi_{i,j}^{(u)}(k)$ gives the probability that the user 'u' is in cell j after k amount of time from the time instant when he/she made transition from somewhere to cell i. However, to predict the location of a user at every k' time steps, we have to estimate the probability $\hat{\phi}_{i,j}^{(u)}(k', s) = P(X_{s+k'}^{(u)} = j | X_0^{(u)} = i, t_{soj} = s)$ i.e., probability that a user is in cell j after k' time given that the current cell is i and user has stayed in cell i for sojourn time $t_{soj} = s$. It can be evaluated as [33]:

$$\hat{\phi}_{i,j}^{(u)}(k', s) = \frac{P(X_{s+k'}^{(u)} = j, t_{soj} = s | X_0^{(u)} = i)}{P(t_{soj} = s | X_0^{(u)} = i)} \tag{15}$$

$$= \frac{h_{i,j}^{(u)}(s+k') + \sum_{m=1}^{z} \sum_{\tau=s+1}^{s+k'} \sigma_{i,m}^{(u)}(\tau) \phi_{m,j}^{(u)}(s+k'-\tau)}{1 - \Lambda_i^{(u)}(s)} \tag{16}$$

Note that for $s = 0 : \hat{\phi}_{i,j}^{(u)}(k', s) = \phi_{i,j}^{(u)}(k)$. We will also leverage steady state distribution of Semi-Markov model to analyze long term cell association of the users. This can help to identify the cells where users spend most of the time and further can be utilized to validate our proposed framework. The steady state distribution of the Semi-Markov i.e., $\zeta^{(u)} = [\zeta_1^{(u)}, \zeta_2^{(u)}, \zeta_3^{(u)}, ..., \zeta_z^{(u)}]$ is given as:

$$\zeta_j^{(u)} = \frac{\pi_j^{(u)} \gamma_j^{(u)}}{\sum_{i=1}^{z} \pi_i^{(u)} \gamma_i^{(u)}} \tag{17}$$

where $[\pi_1^{(u)}, \pi_2^{(u)}, \pi_3^{(u)}, ..., \pi_z^{(u)}]$ is positive solution to following balance equations:

$$\pi_j^{(u)} = \sum_{i=1}^{z} \pi_i^{(u)} p_{i,j}^{(u)}, 1 \leq j \leq z \tag{18}$$

$$\sum_{i=1}^{z} \pi_i^{(u)} = 1 \tag{19}$$

and $\gamma_j^{(u)}, 1 \leq j \leq z$ is the mean sojourn time of user 'u'in cell j. Utilizing the past handover history of user 'u' <time, Cell ID>, Probability transition matrix $\mathbf{P}^{(u)}$ and sojourn time distribution matrix $\mathbf{S}^{(u)}$ are initialized as follows [37]:

$$p_{i,j}^{(u)} = \frac{N_{i,j}^{(u)}}{N_i^{(u)}} \tag{20}$$

and

$$S_{i,j}^{(u)}(k) = \frac{N_{i,j,k}^{(u)}}{N_{i,j}^{(u)}} \tag{21}$$

where $N_{i,j}^{(u)}$ is the number of handovers of user 'u'from cell i to j, $N_{i,j,k}^{(u)}$ is the number of handover of user 'u' from cell i to j with sojourn time less than or equal to k and $N_i^{(u)}$ is the total number of handovers of user 'u' from cell i. Whenever there is a handover from cell i to j, it updates $p_{i,j}^{(u)}$ and $S_{i,j}^{(u)}(k)$ and computes $\psi_{i,j}^{(u)}(k)$. Finally $\phi_{i,j}^{(u)}(k)$ and $\hat{\phi}_{i,j}^{(u)}(k', s)$ are computed. The cell with highest probability is chosen as the predicted future destination i.e., $\max_{j \in \mathbb{N}_i} \hat{\phi}_{i,j}^{(u)}(k', s)$ where $\mathbb{N}_i$ is set of all neighboring cells of cell i. In this way, after every k' time steps, the next HO tuple information for each UE $\{\mathbb{C}_N^u, \mathbb{T}_{HO}^u\}$ is generated wherein $\mathbb{C}_N^u$ is next probable cell of user 'u' at time $\mathbb{T}_{HO}^u$.

C. Future Location Estimation

Let the UE's current location coordinates at time instant k be $l_k^u = (x_k^u, y_k^u)$ and the next cell HO tuple information for each UE be $\{\mathbb{C}_N^u, \mathbb{T}_{HO}^u\}$. Next task is to utilize this information for estimating UE's future location coordinates in next time step $k + k'$. Inspired by observation [38], [39] that nodes in a network usually move around a set of well-visited landmarks with landmark trajectory fairly regular, we utilize past mobility logs of UEs to estimate most probable landmarks visited by each UE in each cell. This information is then utilized to estimate direction of trajectory from current location while distance to be travelled in that direction is estimated using next cell HO time $\mathbb{T}_{HO}$. Let the coordinates of most probable landmark for UE 'u' in next cell $\mathbb{C}_N^u$ be $l_{\mathbb{C}_N^u}^{LM} = (x_{\mathbb{C}_N^u}^{LM}, y_{\mathbb{C}_N^u}^{LM})$ then a unit vector $\hat{u}$ originating from current coordinates in direction of $(x_{\mathbb{C}_N^u}^{LM}, y_{\mathbb{C}_N^u}^{LM})$ is given as:

$$\hat{u} = \frac{[l_{\mathbb{C}_N^u}^{LM} - l_k^u]}{||(l_{\mathbb{C}_N^u}^{LM} - l_k^u)||} \tag{22}$$

where $||.||$ is Euclidian norm operator. The future coordinates at time step $k + k'$ can be estimated as:

$$l_{k+k'}^u = l_k^u + \frac{\sqrt{(x_{\mathbb{C}_N^u}^{LM} - x_k^u)^2 - (y_{\mathbb{C}_N^u}^{LM} - y_k^u)^2}}{T_{HO}^u} * k' * \hat{u} \tag{23}$$

The pseudocode for the next location estimation algorithm is given in Algorithm 1.

D. Proactive Energy Saving Optimization

Given the next probable HO tuple and estimated future location $l_{k+k'}^u$ for all users, we devise ON-OFF sleeping mechanism for SCs for next time step $k + k'$ to minimize network wide energy consumption. The sleeping schedule is ensured to satisfy coverage KPI and QoS requirement of each UE located at its estimated future location $l_{k+k'}^u$ as well as satisfying maximum loading constraint for each BS. The total instantaneous power consumption of a cell can be given by the sum of circuit and the transmit power as [3]:

$$P_c^{\text{total}} = \lambda^c (P_{CT}^{\text{c}} + \eta_c . P_t^{\text{c}}) \tag{24}$$

where P_{CT}^{c} is the constant circuit power which is drawn if BS in cell c is active and is significantly reduced if the BS goes into sleep mode, P_t^{c} is the transmit power of cell c, η_c denotes the load and λ^c is indicator variable that will be 1(0) for ON(OFF) BS in cell c. One way to quantify Energy Savings is to leverage the performance metric criterion of Energy Consumption Ratio (ECR) [40], [41]. This ECR for a cell is defined as the amount of energy consumed in Joules per each bit of information that is reliably transmitted in that cell calculated as:

$$ECR_c = \frac{P}{\sum_{\mathbb{U}_c} \omega_B^u * f(\gamma_u^c)} (Joules/bit) \tag{25}$$

where $f(\gamma_u^c)$ is a function that returns achievable spectral efficiency of user 'u' at a given SINR γ_u^c and ω_B^u is the bandwidth assigned to user 'u'. The $f(\gamma_u^c)$ can be defined to take into account post processing diversity gains such as the ones harnessed by MIMO and/or loss incurred by system specific overheads using $f(\gamma_u^c) := A \log_2(1 + B(\gamma_u^c))$. Here A and B are constants taken as 1 in our simulations studies without loss of generality. The SINR $\hat{\gamma}_u^c$ at an estimated user location $l_{k+k'}^u$ at time step $k+k'$ when associated with a cell c is defined as the ratio of reference signal received power $P_{r,u}^c$ by user 'u' from cell c to the sum of reference signal received power by user 'u' from all cells i such that $\forall i \in \mathbb{C}/c$, and the noise variable κ:

$$\hat{\gamma}_u^c(k + k') = \left[\frac{P_t^c G_u G_u^c \delta \alpha (d_u^c)^{-\beta}}{\kappa + \sum_{\forall i \in \mathbb{C}/c} P_t^i G_u G_u^i \delta \alpha (d_u^i)^{-\beta}} \right]_{k+k'} \tag{26}$$

where P_t^c is the transmit power of cell c, G_u is the gain of user equipment, G_u^c is the gain of transmitter antenna of the cell c as seen by the user 'u', δ is the shadowing observed by the signal, α is the path loss constant, d_u^c represents the distance of estimated user location of 'u' i.e., $l_{k+k'}^u$ from cell c and β is the pathloss exponent. The time subscript on right hand side of (26) and in rest of the paper indicates that all terms enclosed

Algorithm 1 : Future Location Estimation

Input: $l_k^u, \mathbb{C}_N^u, \mathbb{T}_{HO}^u, l_{\mathbb{C}_N^u}^{LM}, SojournTime_{\max}, k, k'$
Output: $l_{k+k'}^u$
for $u \in \mathbb{U}$
  If Sojourn time of u $\geq SojournTime_{\max}$ **OR** no training sample exist for this $\mathbb{C}_N^u$ i.e., $l_{\mathbb{C}_N^u}^{LM} = \{\}$
    $l_{k+k'}^u = l_k^u$
  Else
    $l_{k+k'}^u = l_k^u + \frac{\sqrt{(x_{\mathbb{C}_N^u}^{LM} - x_k^u)^2 - (y_{\mathbb{C}_N^u}^{LM} - y_k^u)^2}}{T_{HO}^u} * k' * \frac{[l_{\mathbb{C}_N^u}^{LM} - l_k^u]}{||(l_{\mathbb{C}_N^u}^{LM} - l_k^u)||}$
  End If
End for

within $[.]_{k+k'}$ are considered for the next time step $k+k'$. In the scope of this paper, it is assumed that shadowing estimate information for the estimated user location is available with normally distributed error. In practical network, Channel Maps building on the Minimization of Drive Test (MDT) reports recently standardized by 3GPP [42] and Channel Quality Indicator reports collected can be utilized to estimate channel gains in estimated locations. This $\hat{\gamma}_u^c(k+k')$ is fully loaded SINR expression and is valid only when all cells are fully utilized. The actual interference from neighboring cells based on their respective loads is utilized as follows to calculate the SINR for data transmission:

$$\hat{\gamma}_u^c(k+k') = \left[\frac{P_t^c G_u G_u^c \delta \alpha (d_u^c)^{-\beta}}{\kappa + \sum_{\forall i \in \mathbb{C}/c} \eta_i P_t^i G_u G_u^i \delta \alpha (d_u^i)^{-\beta}} \right]_{k+k'} \tag{27}$$

where η_i denotes cell load in a cell i at time step $k+k'$. This way of weighting the interference power received from each cell with its current resource utilization yields a certain coupling of the total interference with different cell utilizations. More loaded cells contribute more interference power than less loaded ones [43]. For LTE network, instantaneous cell load can be defined as the ratio of Physical Resource Blocks (PRBs) occupied in cell during a Transmission Time Interval (TTI) and total PRBs available in the cell. This indicator is available as a standard measurement in LTE as "UL/DL total PRB usage". The number of PRBs allocated to each user depends on the QoS that the user requires and achievable SINR. For instance, if the QoS is defined in terms of the required data rate, more PRBs are assigned to a user with higher rate requirement and/or one with lower SINR. The total load of cell c at time step $k+k'$ will be the fraction of the total resources in the cell required to achieve required rate of all users of a cell given as:

$$\eta_c(k+k') = \left[\frac{1}{N_c} \sum_{\mathbb{U}_c} \frac{\hat{\tau}_u}{\omega_B \log_2 1 + \gamma_u^c} \right]_{k+k'} \tag{28}$$

where ω_B is the bandwidth of one resource block, N_c is the total number of resource blocks in cell c, $\hat{\tau}_u$ is the minimum required rate of the user and $\mathbb{U}_c$ is the number of active users connected to a cell c. It is a virtual load as it is allowed to exceed one to give us a clear indication of how overloaded a cell is. The required rate in the numerator is the minimum bit rate required by the user depending upon the QoS requirements of the services and user subscription level. In LTE standard currently there does not exist an exact method to estimate the throughput required by the user. Only historical throughput of user can be estimated after allocation of resources. However, 3GPP standards do define a metric called QoS Class Identifier (QCI). The primary purpose of QCI is to prioritize users based on their required resource type, packet delay susceptibility and packet error loss rate. The definition of desired throughput can build on QCI. In a more robust approach leveraging network analytics, $\hat{\tau}_u$ can be modelled as function of subscriber behavior, subscription level, service request patterns, as well as the applications being used [1]. The set of users connected to cell c is determined by the user association criterion:

$$\mathbb{U}_j := \{\forall u \in \mathbb{U} \mid j = \arg\max_{\forall c \in \mathbb{C}} (P^c_{r,u_{dBm}} + P^c_{CIO\,dB})\} \tag{29}$$

where $P^c_{r,u_{dBm}}$ is the true reference signal power in dBm received by user 'u' from cell c and $P^c_{CIO\,dB}$ is the bias parameter (Cell Individual Offset - CIO). This CIO is primarily used to offset lower transmit power of small cells to transfer more load to them. In case some underutilized cells are turned OFF, remaining cells need to have maximum utilization to cater the transferred load from underutilized cells. However the downside of biasing is that UEs are no longer necessarily connected to the strongest cell. As a result, SINR is bound to be lower with higher CIO values. However, CIO is still a necessary measure to balance the loads. The capacity loss due to drop in SINR can partially be offset if the serving cell has more free PRBs that can be allocated to that user, compared to PRBs in the previous serving cell to satisfy required QoS. This highlights the importance of CIO parameter as a knob to control the tradeoff between network load balancing, CCO and Energy Consumption. It is important to highlight here that in case of ES Optimization with guaranteed minimum QoS requirements, it doesn't make sense to look at throughputs, since the UEs either get exactly the constant bit rate or they are unsatisfied. Hence, more appropriate performance metric to analyze is the number of unsatisfied or dropped users "N_{us}" given as [44]:

$$N_{us}(k+k') = \left[\sum_c \max(0, \sum_{\mathbb{U}_c} 1.(1 - \frac{1}{\eta_c})) \right]_{k+k'} \tag{30}$$

where $\sum_{\mathbb{U}_c} 1.$ sums up to total number of users in cell c while $(1 - \frac{1}{\eta_c})$ is modulation parameter indicating what percentage of users in that cell are unsatisfied. Here η_c by definition from (28) is allowed to exceed 1 to give a clear indication how overloaded a cell is. When $\eta_c = 1$, the inner summation in (30) will be zero meaning all users in cell c are satisfied. When $\eta_c = 2$, the inner summation will be equal to half of the number of users of cell c meaning half of the users are satisfied. Outer summation sums up to total number of unsatisfied users in whole network while max operator is used since the number of unsatisfied users cannot be negative in under loaded cells. The unsatisfied users would not be admitted to enter the system, or they would be dropped if they are already active.

Now we formulate the general energy consumption minimization problem for time step $k+k'$ as (31-33):

$$\min_{\boldsymbol{\lambda_c}, \boldsymbol{P^c_{CIO}}} \sum_{\mathbb{C}} [ECR_c]_{k+k'} \tag{31}$$

The objective is to optimize the parameters $\boldsymbol{\lambda^c}, \boldsymbol{P^c_{CIO}}$ of SCs ($\mathbb{SC}$) such that energy consumption ratio in all cells is minimized while ensuring coverage reliability and satisfaction of user throughput requirements. The first two constraints define the limits for the CIOs and ON/OFF state array respectively. These are the constraints that will determine the size of solution search space. The third constraint is to ensure minimum coverage. Here P^c_{th} is the threshold for the minimum received power for user to be considered covered, $\bar{\omega}$ defines the area coverage probability (a QoS KPI) that operator wants to maintain, and 1(.) denotes indicator function. The fourth

$$\min_{\boldsymbol{\lambda^c}, \boldsymbol{P^c_{CIO}}} \sum_{\mathbb{C}} \left[\frac{\lambda^c (P^c_{CT} + \eta_c . P^c_t)}{\sum_{\mathbb{U}_c} \omega^c_u \log_2 (1 + (\frac{P^c_t G_u G^c_u \delta \alpha (d^c_u)^{-\beta}}{\kappa + \sum_{\forall i \in \mathbb{C}/c} \eta_i P^i_t G_u G^i_u \delta \alpha (d^i_u)^{-\beta}}))} \right]_{k+k'} \tag{32}$$

where $\mathbb{U}_j := \{\forall u \in \mathbb{U} \mid j = \arg\max_{\forall c \in \mathbb{C}} (P^c_{r,u_{dBm}} + P^c_{CIO\,dB})\}$

$$P^c_{CIO.min} \leq P^c_{CIO} \leq P^c_{CIO.max} \forall c \in \mathbb{SC} \tag{33a}$$

$$\lambda^c \in \{0, 1\} \forall c \in \mathbb{SC} \tag{33b}$$

$$\frac{1}{|\mathbb{C}|} \sum_{\mathbb{C}} \frac{1}{|\mathbb{U}_c|} \sum_{\mathbb{U}_c} 1(P^c_{r,u} \geq P^c_{th}) \geq \bar{\omega} \tag{33c}$$

$$\tau_u \geq \hat{\tau}_u \forall u \in \mathbb{U} \tag{33d}$$

$$\eta_c \leq \eta_T \forall c \in \mathbb{C} \tag{33e}$$

constraint ensures each users gets the required minimum bit rate depending upon the QoS requirements of the service and user's subscription level. This is due to the fact that to achieve ECR minimization objective, CIO of the remaining ON SCs may be increased to offload users of switched OFF cells into their coverage umbrella. The consequences are that the received power $P^c_{r,u}$ of offloaded users may become worse, leading to degraded SINR and throughputs. The effect of decreased SINR can be offset by allocating more resources only if the received power by the user is above a certain threshold. Therefore, this fourth constraint ensures that minimum throughput is guaranteed for all users in all cases. However, this can only happen when the number of resources available in a cell are sufficient to meet user requirement, therefore, this constraint is complemented with a constraint on cell load $\eta_c < \eta_T$ (Load Threshold) with $\eta_T \in (0, 1]$. The formulated combinatorial optimisation problem in (32-33) contains both continuous $\boldsymbol{P^c_{CIO}}$ and binary $\boldsymbol{\lambda^c}$ decision variables. It can be identified as a mixed integer non-linear programming problem (MINLP). The inherent coupling of ON/OFF state vector, CIOs and cell loads indicate it is a large scale non convex optimization problem. As we are dealing with two problem parameters per cell whose effects on the optimization function are not independent therefore the complexity is expected to grow exponentially with the number of cells. Hence an exhaustive search for the optimal parameters may not be practical for large size network due to high complexity time search that needs to be done in real time. For a practical scenario with 50 SCs and only CIO as optimization variable with ten possible values available at each SC, we already have 10^{50} possible settings. This is approximately equal to the number of atoms on earth. Therefore in order to solve the formulated ES problem, we utilized Genetic Algorithm (GA) [45]. The reason being it is considered attractive heuristic technique for a multi-variable MINLP problems with a large variable count and enormous search space. Due to its random nature, the genetic algorithm significantly improves chances of finding a global solution especially for highly non-linear objective functions. It is also important to note that the genetic algorithm starts from a random parameter set in the solution space, therefore, does not require a feasible point to start search. Consequently based on estimated network state for time step $k + k'$, AURORA Framework devises optimal ON/OFF state array and CIO values for all the SCs ahead of time such that energy consumption ratio of the whole network is minimized. The ON/OFF state array and CIO values remain fixed from k to k'. As in practical network, SCs need some non-zero time in switching their state therefore the proposed strategy gives ample time of k' duration for SCs to switch to optimal ON/OFF state.

III. Performance Analysis

In this section, we present results for our proposed AURORA Framework. First we analyze the mobility prediction accuracy of the Semi-Markov based model. Then we analyze the potential energy savings resulting from the application of AURORA Framework on HetNets. We have benchmarked its performance against four schemes (i): **Nea**r-Optimal Pe**r**formance Bou**n**d (NARN) wherein it is assumed that AURORA estimates future location and channel estimate at that location with 100% accuracy, (ii): **All** Cell **ON** with **Hom**ogeneous **Net**work Settings (AllOn-HomNet) wherein all cells are ON and no CIO is utilized for small cells, (iii) **All** Cell **On** with **Het**erogeneous **Net**work Settings (AllOn-HetNet) wherein all cells are ON and fixed CIO of 10 dB is utilized for all small cells, (iv) Reactive scheme that is simulated by delaying user location information i.e., Optimization with $\eta_T = 1$ is done based on location information of past one minute.

A. Simulation Settings

We generated typical macro and small cell based network and UE distributions leveraging LTE 3GPP standard compliant [46] network topology simulator in MATLAB. The simulation parameters details are given in Table I. We used wrap around model to simulate interference in an infinitely large network thus avoiding boundary effects. To model realistic networks, UEs were distributed non-uniformly in the coverage area such that a fraction of UEs were clustered around randomly located hotspots in each sector. Monte Carlo style simulation evaluations were used to estimate average performance of the proposed framework. The real challenge here was selection of a mobility trace generation model that realistically represents behavior of actual cellular network users. Several such models

TABLE I
NETWORK SCENARIO SETTINGS

System Parameters	Values
Number of Macro Base Stations	7 with 3 Sectors per Base Station
Small Cells per Sector	5
Number of UEs	Mobile: 84, Stationary: 336
LTE System Parameters	Frequency = 2 GHz, Bandwidth = 10 MHz
Macro Cell Tx Parameters	Tx Power = 46 dBm, Tilt = 10^{2^0}
Small Cell Tx Parameters	Tx Power = 30 dBm, CIO = 0 to 10 dB
Base Station Heights	Macro BS = 25m, Small BS = 10m
Area Coverage Probability	100%
Total Simulation Duration	1 hour

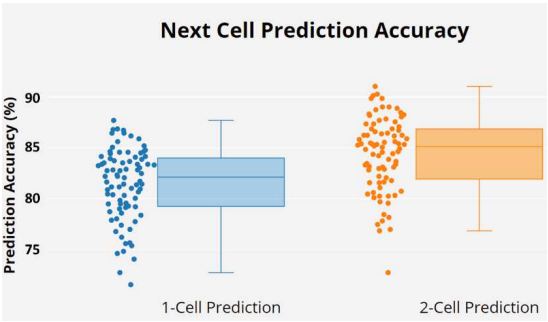


Fig. 3. Next Cell Prediction Accuracy

have been proposed recently in literature such as SLAW, SMOOTH, Truncated Levy Walk etc., [47]. Based on an extensive analysis of pros and cons of these models, we chose SLAW (Self-similar Least Action Walk) [48] mobility model. Contrary to the conventional random walk models where movement at each instant is completely random, chosen randomly from set of allowed speed and angles, SLAW has been shown to be a highly realistic mobility model. It exhibits all the characteristics of real world human mobility i.e., (i) ***truncated power-law flights and pause-times:*** the lengths of human flights which are defined to be straight line trips without directional change or pause have a truncated power-law distribution (ii) ***heterogeneously bounded mobility areas:*** people mostly move only within their own confined areas of mobility and different people may have widely different mobility areas (iii) ***truncated power-law inter- contact times:*** the times elapsed between two successive contacts of the same persons follows truncated power law distribution and (iv) ***fractal waypoints:*** people are always more attracted to more popular places. Therefore, the accuracy of AURORA Framework tested using mobility traces generated by SLAW is very likely to represent its true performance in real network. The SLAW mobility model was utilized to generate HO traces of 84 mobile users for one week. Out of which, traces for first six days were utilized to build and train Semi-Markov mobility model for each of the 84 UEs. Moreover, additional 336 stationary UEs (80% of total UEs [49]) were deployed to generate additional loading on the network. For Traffic Demand, we considered two scenarios (i) ***Low Traffic Demand*** comprising of five different uniformly distributed UE traffic requirement profiles corresponding to 24 kbps (voice), 56 kbps (Text Browsing), 128 kbps (Image Browsing), 512 kbps (FTP) and 1024 kbps (video) desired throughputs, (ii) ***High Traffic Demand*** wherein all UEs are video users. Without loss of generality and keeping operational complexity in mind, the prediction interval k' was set as 1 minute in our simulation study.

B. Mobility Prediction Accuracy

For benchmarking prediction accuracy of the Semi-Markov based model trained on six days training data, we utilized

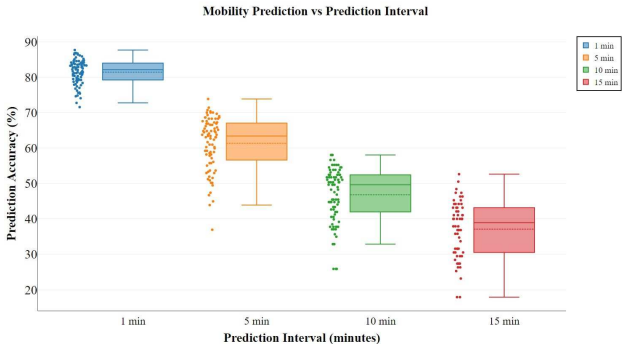


Fig. 4. Effect of Prediction Interval on Next Cell Prediction Accuracy

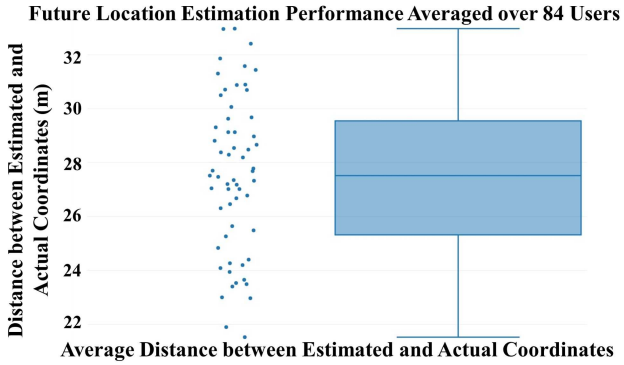


Fig. 5. Future Location Coordinates Estimation Performance

(13) and (16) to predict serving cells of all UEs for the next whole day after every k' time step. At each time interval k, when the predicted future cell in next time interval k' is same as actual future cell than score given is 1 otherwise 0. Accuracy is then calculated by summing scores for all time instants and divided by number of observations. The next cell prediction accuracy results are given in Fig. 3. Accordingly, maximum prediction accuracy of 87.70% was achieved having mean value of 81.46% when choosing the top most probable cell among all future next cell candidates (1-Cell Prediction). The predictor performs exceptionally well since prediction interval is only one minute. This high prediction accuracy is in line with our recent published study [28] on benchmarking prediction accuracy of Semi-Markov based mobility prediction model using Real HO measurements collected from live LTE network. This prediction can be enhanced further by decreasing k' interval length. Fig. 4 shows mean prediction accuracy (denoted by dotted lines) monotonically decreases with the increase in k' interval length. We could not decrease prediction interval to less than 1 minute as with computational resources available for this study Genetic Algorithm needed at least this minimum amount of time to find a feasible solution. However, it is anticipated that if more powerful computational resources are leveraged to reduce the convergence time of Genetic algorithm, better mobility prediction accuracy may be achieved. We also analyzed the effect of choosing the two top most probable future next cell candidates (2-Cell Prediction) instead of one. The prediction accuracy got a little boost with mean value reaching up-to 84.39%. However this gain is not that significant given it already has very high accuracy. Next, based on next cell HO tuple information for each UE $\{\mathbb{C}_N^u, \mathbb{T}_{HO}^u\}$, future location coordinates were estimated using

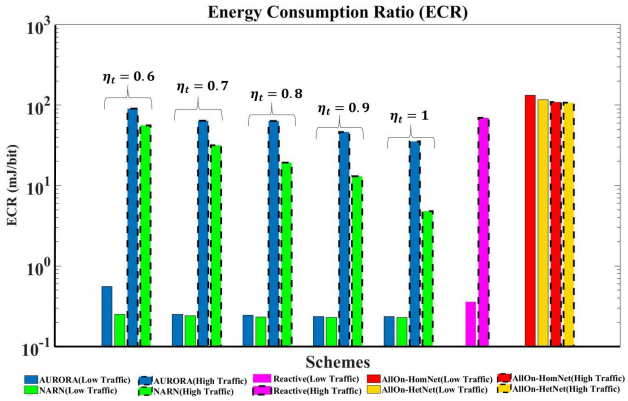


Fig. 6. Energy Consumption Ratio (ECR)

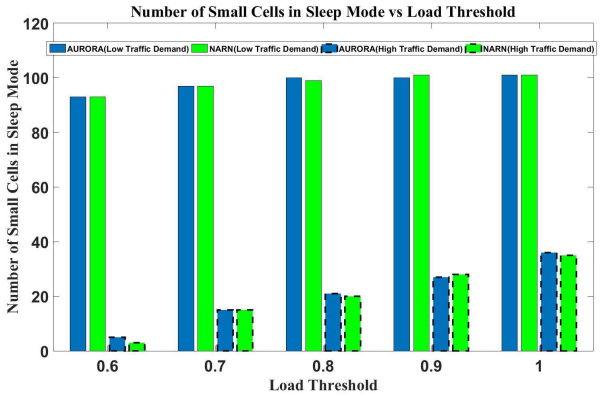


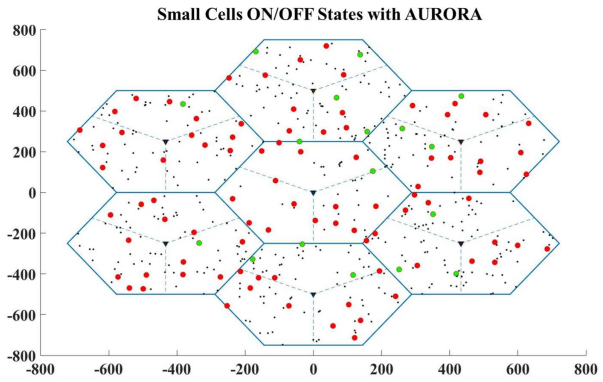
Fig. 7. Number of SCs put to Sleep Mode vs Load Threshold

Algorithm 1 for all UEs for one hour simulation duration after every k' time steps. The average estimation performance is illustrated in Fig. 5 according to which maximum distance error between estimated and actual coordinates was around 33 meters having mean value of around 27.5 meters. The location estimation algorithm performed exceptionally well. One particular reason for high accuracy is that SLAW model is for pedestrian users. Therefore, location of user changes slowly as function of time and thus remains relatively more predictable. With high speed, accuracy is expected to degrade, but then knowledge of street/road layout can be exploited to maintain accuracy. However, this is beyond scope of this paper and will be subject of future study. An interesting observation stemming from the symmetric shape of Box Plot and absence of outliers suggest that normal distribution can be good approximation for the average location estimation error distribution.

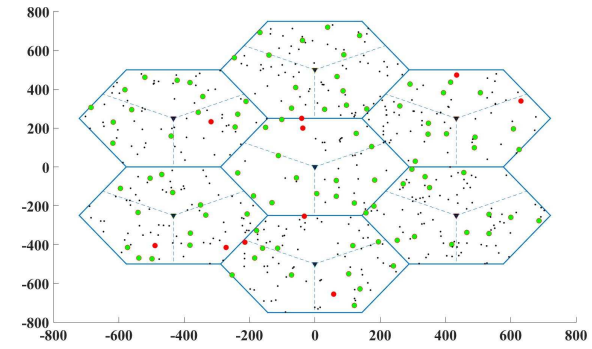
C. Quantifying Energy Saving Potential of AURORA Framework

The Energy Consumption Ratio (ECR) of AURORA and NARN for Low and High Traffic Demands with varying values of Load thresholds η_T along with that of AllOn-HomNet, AllOn-HetNet and state of the art Reactive schemes averaged over 1 hour duration is visualized in Fig. 6. Note that for visualizing ECR ranges for both Traffic Classes in same figure, the y-axis has been plotted in logarithmic scale. The load threshold range is [0.6, 1] since below 0.6 there was no feasible point returned by the P-ES optimization algorithm (33). It is observed that ECR values are higher for high traffic demand scenario as more number of SCs need to be switched ON to cater high load. Moreover AURORA exhibit a linearly decreasing trend with increasing values of η_T. It is significantly much less than the conventional AllOn schemes for all load threshold values. The reason being that for AllOn schemes, all cells are ON at all times that increases energy consumption which is bound to further escalate with densification. At lower η_T values, ECR for AURORA is higher since smaller η_T value compels the AURORA to keep ON larger number of underutilized SCs. For instance at $\eta_T = 0.6$, AURORA switches ON next small cell as soon as the utilization of current ON small cells reach 60%. Thus, on average, large number of SCs will be turned ON for smaller η_T values thus increasing energy consumption. Moreover, with large number of SCs turned ON, there is higher chance that location estimation inaccuracy results in turning ON SCs with very low or no load (i.e., very high ECR - Joules/bit). On the other hand, larger values of η_T enables AURORA to switch OFF large number of SCs. For instance at $\eta_T = 1$, AURORA will switch ON next SC only when the utilization of current ON SCs reaches 100%. As a result ECR is expected to decrease and same trend is observed for NARN. It is interesting to observe that on one hand with increasing value of η_T, less number of SCs are turned ON therefore there is less chance of any turned ON SCs with very low or no load. On the other hand, with increasing η_T values, AURORA switches ON smallest possible number of SCs and all of them almost fully utilized with very few resources to spare. As a result inaccuracy in location estimation will result in increased risk of blocking of the UEs (hence increased number of unsatisfied users $-$ see Fig. 9) thereby negatively affecting QoS. However, as number of fully utilized SCs is a more dominant factor in determining overall ECR as compared to slight increase in the number of unsatisfied users, therefore overall ECR reduces. The comparison of AURORA with Reactive scheme shows that ECR for Reactive scheme is higher as compared to AURORA. This is because in Reactive scheme, due to delayed user location information outdated configuration settings that are suboptimal for current instant are applied to the network. This increases the percentage of unsatisfied users (on average 1.85% with AURORA at $\eta_T = 1$ while 4% with Reactive scheme at high traffic load) and hence higher ECR. Moreover, ECR for AllOn-HomNet is slightly higher as compared to AllOn-HetNet. This is because higher CIO values used in AllOn-HetNet compels SCs to be more utilized and hence reduced ECR as compared to AllOn-HomNet scheme.

Fig. 7 shows the average number of SCs put to sleep mode with AURORA and NARN with varying values of η_T for low and high traffic demand. It can be seen that less number of SCs can be put to sleep mode for meeting needs of high traffic demand. The number of SCs put to sleep mode continue to increase with η_T. This is because with increasing values of η_T, a SC is utilized more before turning ON next SC or in other words more SCs are put to sleep mode at higher values of η_T. Since load coupled interference also increases with η_T therefore optimization algorithm returns such an Optimization Parameters Configuration (OPC) i.e., $\boldsymbol{\lambda^c}, \boldsymbol{P^c_{CIO}}$



(a) Low Traffic Demand



(b) High Traffic Demand

Fig. 8. : Snapshot for Small Cells (ON/OFF) States by AURORA for (a) Low Traffic Demand (b) High Traffic Demand. Green (Red) circles indicate ON(OFF) SCs and UEs are illustrated by black dots.

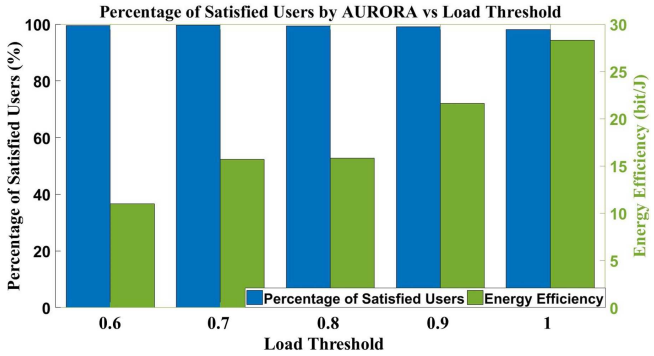


Fig. 9. Percentage of Satisfied Users vs Load Threshold for High Traffic Demand

that minimizes overall energy consumption ratio. A snapshot for the SCs states with AURORA for low and high traffic scenarios at same time instants are shown in Fig. 8. It can be observed that for high traffic demand, majority of the SCs are turned ON. For space limitation, results in all subsequent figures correspond to high traffic demand scenario only that follow same trend as that observed with low traffic demand. The average percentage of satisfied users under AURORA framework vs Load Threshold η_T for high traffic demand scenario is visualized in Fig. 9 on left y-axis while Energy Efficiency $(1/ECR)$ is plotted on right y-axis. It can be observed at low η_T values, plenty of free resources are available in relatively more number of available BSs hence more users are served with enough resources to meet their minimum QoS requirements. Even with location estimation inaccuracies, the

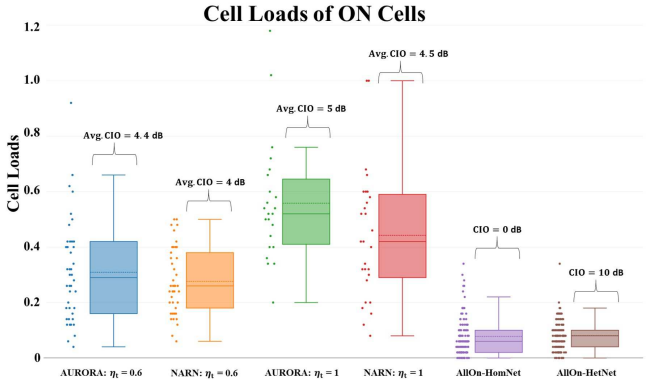


Fig. 10. Cell Loads of ON Cells for High Traffic Demand

UEs will still have better chance to get enough resources and be satisfied. However, more SCs are turned ON at low η_T with more chance of being underutilized and hence lower Energy Efficiency. As η_T value becomes higher and approaches 1, AURORA returns such an OPC $\boldsymbol{\lambda^c}, \boldsymbol{P^c_{CIO}}$ that results in smallest possible number of switched ON SCs and all of them almost fully utilized with very few resources to spare. Hence a slight location estimation inaccuracy can result in increased risk of blocking and hence decrease in number of satisfied users. Contrary to that, fewer cells turned ON with more utilization improve energy efficiency of the network. It is interesting to observe that for high traffic demand scenario even at $\eta_T = 1$, percentage of satisfied users is above 98%. The cell loads of ON cells achievable with AURORA and NARN with $\eta_T = 0.6$ and 1 alongside with AllOn schemes for high traffic demand is plotted in Fig. 10.

It is evident from the figure that in case of AllOn-HomNet and AllOn-HetNet, since all cells are kept ON, therefore most of the cells are underutilized with mean utilization of 7.74% and 8% in AllOn-HomNet and AllOn-HetNet respectively. This results in higher ECR (see Fig. 6). With AURORA and NARN, at lower value of η_T i.e., 0.6, some SCs are switched OFF and thus utilization of remaining ON cells relatively increases with mean utilization of 30.9% and 27.6% respectively. At higher value of η_T i.e., 1, large no. of SCs are switched OFF and the few ones which are ON, are relatively more utilized with mean utilization of 55.8% and 44.2% respectively. The average CIO values are indicated on top of each boxplot. It is observed that at higher η_T value of 1 as compared to lower value of 0.6, on average, relatively larger CIO values have been leveraged. This is because when fewer cells are switched ON, CIO values of ON SCs are boosted up to serve the users of OFF cells. In this way CIOs complements the Proactive Energy Consumption Optimization by serving as a guiding parameter in directing users to suitable cells such that overall ECR reduces while satisfying QoS requirements. The results for average downlink SINR for AURORA and NARN with $\eta_T = 0.6$ and 1 along with the AllOn-HomNet and AllOn-HetNet for High Traffic Demand Scenario is shown in CDF plot in Fig. 11. It can be observed that at higher value of η_T i.e., 1, load coupled interference from neighboring BSs is very high. Therefore SINR is negatively affected for AURORA and NARN as compared to AllOn-HomNet and AllOn-HetNet. As a matter of fact, when CIOs are leveraged, degraded SINR is natural outcome. However it does not mean a degraded system

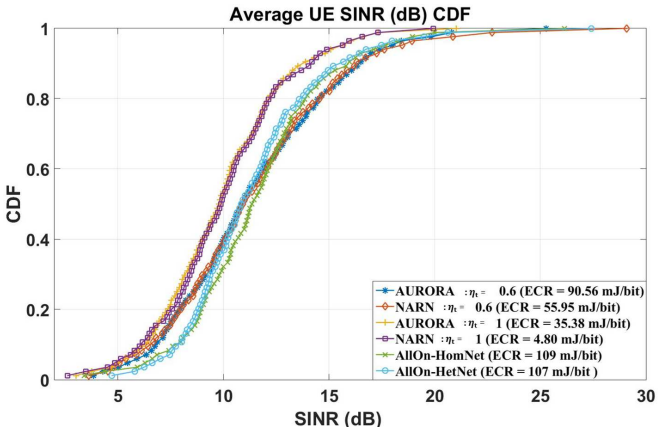


Fig. 11. Average UE SINR CDF for High Traffic Demand

wide performance as long as loss in throughput caused by lower logarithmic SINR term is offset by increased number of PRBs allocable to users. This is how it strives to guarantees their minimum QoS requirements as shown in Fig. 9. At smaller η_T value of 0.6, more number of SCs are turned ON with relatively less load. This reduces overall interference floor in the network and hence SINR improves that is higher than that achievable at η_T value of 1. For AllOn-HomNet and AllOn-HetNet schemes, all SCs are ON, highly underutilized and hence higher SINR. However, it is worth noting that this gain in SINR comes at cost of higher energy consumption i.e., for AllOn-HomNet and AllOn-Hetnet, ECR is 109 mJ/bit and 107 mJ/bit respectively that is much higher as compared to AURORA which is around 36 mJ/bit achievable at $\eta_T = 1$.

The average long term cell occupancy probability of the users computed through (17-19) is shown in Fig. 12(a) according to which users spend most of their time in Macro cells 5, 1, 19, 20 and 21 (denoted by yellow stars). This information can be utilized for validation of the proposed AURORA Framework. The average percentage of ON Small Cells with AURORA for one hour simulation duration is shown in Fig. 12(b). As is evident, more number of SCs were turned ON in Macro cells 9, 20, 5, 19 and 1 (denoted by yellow stars). Hence on average AURORA kept more number of SCs switched ON in cells where users had higher sojourn time. Few discrepancies observed such as with Macrocell 21 can be attributed to the location estimation inaccuracies as well as rate requirement of UEs in those cells i.e., even with higher cell occupancy probability of users in a particular macrocell, if cumulative rate requirement of UEs is low than SCs in that macrocell will remain switched OFF most of the time. For higher traffic demand scenario, average percentage of ON Small Cells with AURORA is shown in figure 12(c). As more number of SCs were turned ON to cope with high traffic demand therefore the plot in Fig. 12(c) is relatively more greenish as compared to that in Fig. 12(b).

D. Quantifying Effect of Mobility Prediction Model Inaccuracy on Potential Energy Saving

The potential energy savings resulting from the application of AURORA Framework can be quantified by computing Energy Reduction Gain (ERG) [40], [41] performance metric given as:

$$ERG = \left(\frac{ECR_{Benchmark} - ECR_{AURORA}}{ECR_{Benchmark}}\right) \times 100\% \qquad (34)$$

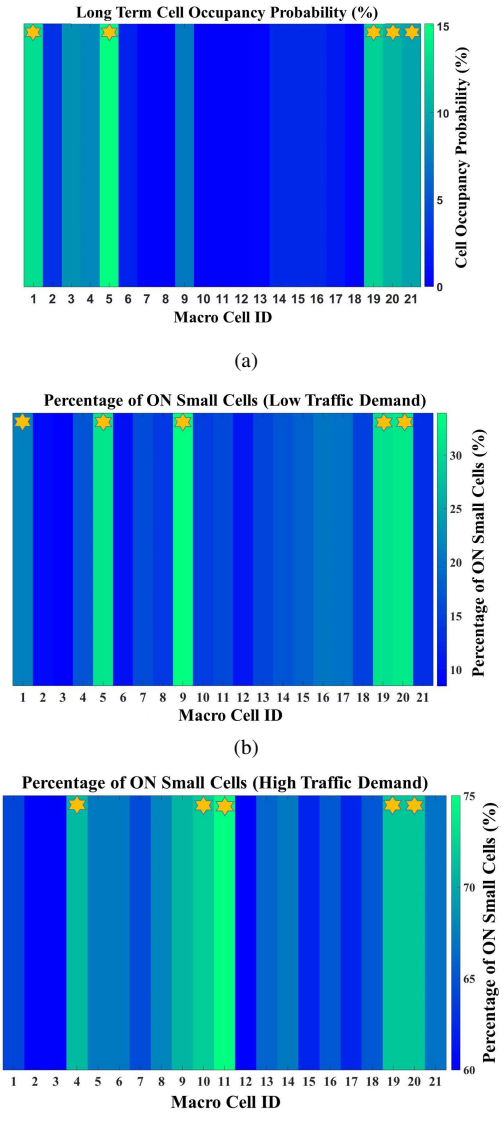


Fig. 12. (a) Long term Cell Occupancy probability (b) Percentage of ON Small Cells at Low Traffic Demand (c) Percentage of ON Small Cells at High Traffic Demand.

It is logical to anticipate that the energy saving gain of AURORA i.e., ERG will depend on the accuracy of the underlying mobility prediction model. In this section we analyze this dependence by varying the underlying user mobility model such that it includes varying degree of randomness and hence predictability. To vary the degree of randomness in the mobility traces, the two key parameters of SLAW mobility model namely variance in pause times and percentage of random waypoints were changed from default values suggested in [48], (and used for results in figures 3-12) to larger values to increase randomness in the mobility trajectory of the UEs. Four set of gradually increasing initialization parameters were used that resulted in increasing randomness in user mobility. Our prediction model trained on these four set of traces exhibited average prediction accuracy of 85%, 75% , 65% and 55%. The average ERG of AURORA for these varying values of Prediction Accuracy against AllOn-HomNet and AllOn-HetNet schemes averaged over 1 hour duration for high traffic

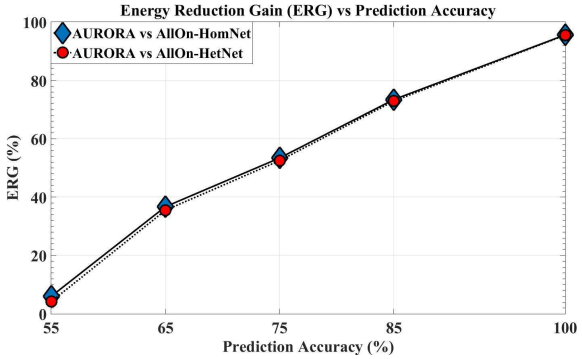


Fig. 13. Energy Reduction Gain vs Prediction Accuracy

demand scenario is plotted in Fig. 13. It is observed that as expected the gain of AURORA decreases with decrease in prediction accuracy. However, it is noteworthy that as long as mobility is predictable with 55% or higher accuracy, AURORA continues to yield Energy Reduction Gain. Given that typical human mobility features 93% predictability when averaged over a large real user sample space [27], AURORA is a promising approach. However, human mobility is bound to have some randomness that translates to prediction inaccuracy. The high frequency periodic update aspect of the future location probabilities is one of the possible ways to cope with the prediction inaccuracies as the effect of the prediction inaccuracy is only limited to the prediction interval. Another way is to make it adaptive so that AURORA continuously analyzes its performance and falls back to conventional AllOn scheme when prediction accuracy drops below 55%. Moreover, selecting top-2 probable locations as shown in Fig. 3, can also be chosen as a strategy to improve the prediction accuracy, albeit at cost of reduced ERG.

IV. Conclusions

This paper has proposed a novel spatiotemporal mobility prediction aware proactive sleep-mode based energy saving optimization algorithm for cracking the future 5G ultra-dense HetNets puzzle. The proposed AURORA framework employs innovative concept of estimating future user locations and leverage that to estimate future cell loads. It then devises energy saving optimization problem for the estimated future network scenario. The majority of the conventional reactive style approaches are expected to solve the formulated energy saving problem dynamically in real-time as network conditions change. However this is close to impossible even when substantial computing power is available. Contrary to that, the innovative proposed approach enables state-of-the-art heuristic techniques like GA to find practically good solutions to the formulated optimization problem predictively ahead of time. This can be enabler for meeting 5G ambitious latency and QoS requirements. Moreover, AURORA framework considers the interplay among the three intertwined SON functions (ES, CCO and LB) due to the overlap among their primary optimization parameters. Therefore it employs co-design approach wherein the joint optimization of ON/OFF States and CIO values for SCs does not conflict with CCO and LB objectives. Extensive simulations employing realistic SLAW mobility model indicate that, in best case, AURORA can achieve energy reduction gain of about 68% for high traffic demand scenario in ultra-dense HetNets as compared to Always On approach. Comparative performance analysis with near-optimal performance bound indicate satisfactory robustness of the proposed AURORA framework towards location estimation accuracies. For future works, we will investigate incorporation of user specific CIOs by considering mobility behavior and QoS requirements of the UEs. We will also investigate incorporating the backhaul constraint implicitly by assigning maximum load threshold to the cells depending upon available backhaul. Another promising research direction is to devise energy aware association scheme and use it in conjunction with the energy saving optimization problem.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant Numbers 1619346, 1559483, 1718956 and 1730650. The statements made herein are solely the responsibility of the authors. For more details, please visit: http://www.ai4networks.com.

References

[1] A. Imran and A. Zoha, "Challenges in 5G: how to empower SON with big data for enabling 5G," *IEEE Network*, vol. 28, no. 6, pp. 27–33, nov 2014.

[2] I. Ashraf, F. Boccardi, and L. Ho, "SLEEP mode techniques for small cell deployments," *IEEE Communications Magazine*, vol. 49, no. 8, pp. 72–79, aug 2011.

[3] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Communications*, vol. 18, no. 5, pp. 40–49, oct 2011.

[4] S. Buzzi, C.-L. I, T. E. Klein, H. V. Poor, C. Yang, and A. Zappone, "A Survey of Energy-Efficient Techniques for 5G Networks and Challenges Ahead," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 697–709, apr 2016.

[5] M. Ajmone Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal Energy Savings in Cellular Access Networks," in *2009 IEEE International Conference on Communications Workshops*. IEEE, jun 2009, pp. 1–5.

[6] R. Litjens and L. Jorguseski, "Potential of energy-oriented network optimisation: Switching off over-capacity in off-peak hours," in *21st Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE, sep 2010, pp. 1660–1664.

[7] Z. Niu, "TANGO: traffic-aware network planning and green operation," *IEEE Wireless Communications*, vol. 18, no. 5, pp. 25–29, oct 2011.

[8] 3GPP, "LTE; Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-configuring and self-optimizing network (SON) use cases and solutions (3GPP TR 36.902 version 9.2.0 Release 9)," Tech. Rep., 2010.

[9] F. Z. Kaddour, E. Vivier, L. Mroueh, M. Pischella, and P. Martins, "Green Opportunistic and Efficient Resource Block Allocation Algorithm for LTE Uplink Networks," pp. 4537–4550, 2015.

[10] S. Wu, Z. Zeng, and H. Xia, "Load-Aware Energy Efficiency Optimization in Dense Small Cell Networks," *IEEE Communications Letters*, vol. 21, no. 2, pp. 366–369, feb 2017.

[11] R. Tao, J. Zhang, and X. Chu, "An Energy Saving Small Cell Sleeping Mechanism with Cell Expansion in Heterogeneous Networks," in *IEEE 83rd Vehicular Technology Conference (VTC Spring)*, May 2016, pp. 1–5.

[12] Y. Qu, Y. Chang, Y. Sun, and D. Yang, "Equilibrated Activating Strategy with Small Cell for Energy Saving in Heterogeneous Network," in *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, Sep 2014, pp. 1–6.

[13] A. Ebrahim and E. Alsusa, "Interference and Resource Management Through Sleep Mode Selection in Heterogeneous Networks," *IEEE Transactions on Communications*, pp. 1–1, 2016.

[14] L.-P. Tung, L.-C. Wang, and K.-S. Chen, "An interference-aware small cell on/off mechanism in hyper dense small cell networks," in *International Conference on Computing, Networking and Communications (ICNC)*, Jan 2017, pp. 767–771.

[15] Q. Wang and J. Zheng, "A Distributed base station On/Off Control Mechanism for energy efficiency of small cell networks," in *2015 IEEE International Conference on Communications (ICC)*. IEEE, jun 2015, pp. 3317–3322.

[16] I. L. C. Araujo and A. Klautau, "Traffic-aware sleep mode algorithm for 5G networks," in *2015 International Workshop on Telecommunications (IWT)*. IEEE, jun 2015, pp. 1–5.

[17] S. Samarakoon, M. Bennis, W. Saad, and M. Latva-aho, "Opportunistic sleep mode strategies in wireless small cell networks," in *IEEE International Conference on Communications (ICC)*, June 2014, pp. 2707–2712.

[18] B. Partov, D. J. Leith, and R. Razavi, "Energy-aware configuration of small cell networks," in *IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, Sep 2014, pp. 1403–1408.

[19] Y. Liu, H. Tian, and G. Nie, "QoS-Aware Distributed Cell Sleep Algorithm for OFDMA Small Cell Networks," in *IEEE 82nd Vehicular Technology Conference (VTC2015-Fall)*, Sep 2015, pp. 1–5.

[20] S. Zhang, N. Zhang, S. Zhou, J. Gong, Z. Niu, and X. Shen, "Energy-Aware Traffic Offloading for Green Heterogeneous Networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1116–1129, may 2016.

[21] Y. Sun, Y. Chang, S. Song, and D. Yang, "An energy-efficiency aware sleeping strategy for dense multi-tier HetNets," in *IEEE Globecom Workshops (GC Wkshps)*, Dec 2014, pp. 1180–1185.

[22] Z. Li, D. Grace, and P. Mitchell, "Traffic-Aware Cell Management for Green Ultradense Small-Cell Networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2600–2614, mar 2017.

[23] E. Oh, K. Son, and B. Krishnamachari, "Dynamic Base Station Switching-On/Off Strategies for Green Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2126–2136, may 2013.

[24] S. Navaratnarajah, A. Saeed, M. Dianati, and M. Imran, "Energy efficiency in heterogeneous wireless access networks," *IEEE Wireless Communications*, vol. 20, no. 5, pp. 37–43, oct 2013.

[25] K. Samdanis, P. Rost, A. Maeder, M. Meo, and C. Verikoukis, *Green Communications: Principles, Concepts and Practice*. Wiley & Sons, Ltd., 2015.

[26] H. Y. Lateef, A. Imran, and A. Abu-dayya, "A framework for classification of Self-Organising network conflicts and coordination algorithms," in *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, sep 2013, pp. 2898–2903.

[27] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of Predictability in Human Mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[28] H. Farooq and A. Imran, "Spatiotemporal Mobility Prediction in Proactive Self-Organizing Cellular Networks," *IEEE Communications Letters*, vol. 21, no. 2, pp. 370–373, feb 2017.

[29] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez, "Next Place Prediction Using Mobility Markov Chains," in *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, ser. MPM '12. New York, NY, USA: ACM, 2012, pp. 3:1—-3:6.

[30] D. Katsaros and Y. Manolopoulos, "Prediction in wireless networks by Markov chains," *Wireless Communications, IEEE*, vol. 16, no. 2, pp. 56–64, apr 2009.

[31] N. A. Amirrudin, S. H. S. Ariffin, N. Malik, and N. E. Ghazali, "User's mobility history-based mobility prediction in LTE femtocells network," in *RF and Microwave Conference (RFM), 2013 IEEE International*, dec 2013, pp. 105–110.

[32] X. Zhou, Z. Zhao, R. Li, Y. Zhou, J. Palicot, and H. Zhang, "Human Mobility Patterns in Cellular Networks," *IEEE Communications Letters*, vol. 17, no. 10, pp. 1877–1880, Oct 2013.

[33] J.-K. Lee and J. C. Hou, "Modeling Steady-state and Transient Behaviors of User Mobility: Formulation, Analysis, and Application," in *Proceedings of the 7th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, ser. MobiHoc '06. New York, NY, USA: ACM, 2006, pp. 85–96.

[34] H. Abu-Ghazaleh and A. S. Alfa, "Application of Mobility Prediction in Wireless Networks Using Markov Renewal Theory," *Vehicular Technology, IEEE Transactions on*, vol. 59, no. 2, pp. 788–802, feb 2010.

[35] I. Schumm, "Lessons Learned from Germany's 2001-2006 Labor Market Reforms," Ph.D. dissertation, Universotu of Wurzburg, 2009.

[36] G. Corradi, J. Janssen, and R. Manca, "Numerical Treatment of Homogeneous Semi-Markov Processes in Transient Cases-a Straightforward Approach," *Methodology And Computing In Applied Probability*, vol. 6, no. 2, pp. 233–246, 2004.

[37] V. Barbu and N. Limnios, "Nonparametric Estimation for Failure Rate Functions of Discrete Time semi-Markov Processes," in *Probability, Statistics and Modelling in Public Health*, M. Nikulin, D. Commenges, and C. Huber, Eds. Springer US, 2006, pp. 53–72.

[38] J. Ghosh, S. J. Philip, and C. Qiao, "Sociological Orbit Aware Location Approximation and Routing (Solar) in DTN," State Univ. of New York at Buffalo, Tech. Rep., 2005.

[39] Q. Yuan, I. Cardei, and J. Wu, "An Efficient Prediction-Based Routing in Disruption-Tolerant Networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 1, pp. 19–31, jan 2012.

[40] F. Cao and Z. Fan, "The tradeoff between energy efficiency and system performance of femtocell deployment," in *7th International Symposium on Wireless Communication Systems*, Sep 2010, pp. 315–319.

[41] B. Badic, T. O'Farrrell, P. Loskot, and J. He, "Energy Efficient Radio Access Architectures for Green Radio: Large versus Small Cell Size Deployment," in *IEEE 70th Vehicular Technology Conference Fall*, Sep 2009, pp. 1–5.

[42] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on Minimization of drive- tests in Next Generation Networks; (Release 9)," Tech. Rep., 2009.

[43] A. J. Fehske, I. Viering, J. Voigt, C. Sartori, S. Redana, and G. P. Fettweis, "Small-Cell Self-Organizing Wireless Networks," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 334–350, mar 2014.

[44] I. Viering, M. Dottling, and A. Lobinger, "A Mathematical Perspective of Self-Optimizing Wireless Networks," in *2009 IEEE International Conference on Communications*. IEEE, jun 2009, pp. 1–6.

[45] S. Luke, *Essentials of Metaheuristics (Second Edition)*. Lulu, 2013.

[46] 3GPP, "3rd Generation Partnership Project; Physical layer aspects for evolved universal terrestrial radio access (e-utra)," TR 25.814 V7.1.0 Release 7, Tech. Rep., 2006.

[47] M. Gorawski and K. Grochla, *Review of Mobility Models for Performance Evaluation of Wireless Networks*. Cham: Springer International Publishing, 2014, pp. 567–577.

[48] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "SLAW: A New Mobility Model for Human Walks," in *IEEE INFOCOM 2009 - The 28th Conference on Computer Communications*. IEEE, apr 2009, pp. 855–863.

[49] Huawei, "WhitePaper: Five Trends to Small Cell 2020," Barcelona, Tech. Rep.

Hasan Farooq (GSM'14) received his B.Sc. degree in Electrical Engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2009 and the M.Sc. by Research degree in Information Technology from Universiti Teknologi PETRONAS, Malaysia in 2014 wherein his research focused on developing ad-hoc routing protocols for smart grids. Currently he is pursuing the Ph.D. degree in Electrical and Computer Engineering at the University of Oklahoma, USA. His research area is Big Data empowered Proactive Self-Organizing Cellular Networks focusing on Intelligent Proactive Self-Optimization and Self-Healing in HetNets utilizing dexterous combination of machine learning tools, classical optimization techniques, stochastic analysis and data analytics. He has been involved in multinational QSON project on Self Organizing Cellular Networks (SON) and is currently contributing to two NSF funded projects on 5G SON. He is recipient of Internet Society (ISOC) First Time Fellowship Award towards Internet Engineering Task Force (IETF) 86th Meeting held in USA, 2013 and winner of the IEEE Young Professional Green ICT Idea Competition 2017.

Ahmad Asghar (S'17) received his B.Sc. degree in Electronics Engineering from Ghulam Ishaq Khan Institute of Science and Technology, Pakistan, in 2010 and the M.Sc. degree in Electrical Engineering from Lahore University of Management and Technology, Pakistan in 2014. Currently he is pursuing the Ph.D. degree in Electrical and Computer Engineering at the University of Oklahoma, USA as well as contributing to multiple NSF funded studies on 5th Generation Cellular Networks. His research work includes studies on Self-Healing and SelfCoordination of Self-Organizing Functions in Future Big-Data Empowered Cellular Networks using analytical and machine learning tools.

Ali Imran (M'15)) received the B.Sc. degree in Electrical Engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2005 and the M.Sc. degree (with distinction) in mobile and satellite communications and the Ph.D. degree from the University of Surrey, Guildford, U.K., in 2007 and 2011, respectively. He is an Assistant Professor in Telecommunications at the University of Oklahoma, Tulsa, OK, USA where he is the founding director of Artificial Intelligence (AI) for Networks Laboratory (AI4Networks) Research Center and TurboRAN 5G Testbed. He has been leading several multinational projects on Self Organizing Cellular Networks such as QSON, for which he has secured research grants of over $3 million in last four years as lead principal investigator. Currently he is leading four NSF funded Projects on 5G amounting to over $2.2 million. He has authored over 60 peer-reviewed articles and presented a number of tutorials at international forums, such as the IEEE International Conference on Communications, the IEEE Wireless Communications and Networking Conference, the European Wireless Conference, and the International Conference on Cognitive Radio Oriented Wireless Networks, on his topics of interest. His research interests include self-organizing networks, radio resource management, and big-data analytics. Dr. Imran is an Associate Fellow of the Higher Education Academy (AFHEA), U.K., and a member of the Advisory Board to the Special Technical Community on Big Data of the IEEE Computer Society.